AFHRL-TP-87-6

(12)

AIR FORCE HUMAN RESOURCES LABORATORY

APPROPRIATENESS MEASUREMENT

Fritz Drasgow
Michael V. Levine
Mary E. McLaughlin

Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432

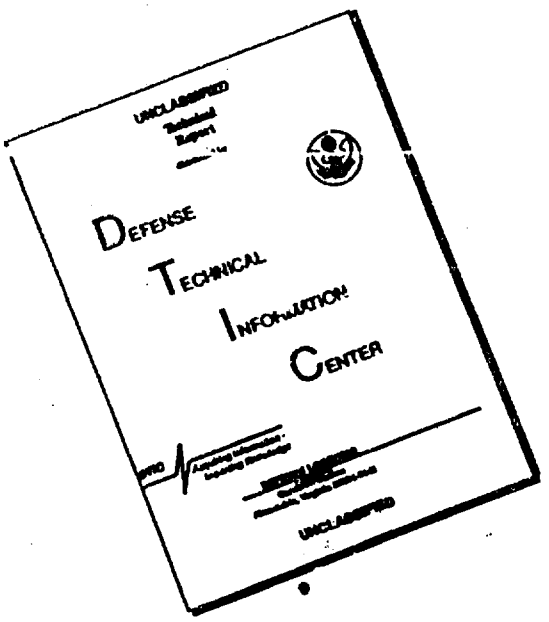
James A. Earles

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

August 1987
Final Technical Paper for Period July 1984 - December 1985

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

RONALD L. KERCHNER, Colonel, USAF
Chief, Manpower and Personnel Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

ADA184185

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-87-6
6a. NAME OF PERFORMING ORGANIZATION Universal Energy Systems, Inc. 6b. OFFICE SYMBOL (*If applicable*)	7a. NAME OF MONITORING ORGANIZATION Manpower and Personnel Division
6c. ADDRESS (*City, State, and ZIP Code*) 4401 Dayton-Xenia Road Dayton, Ohio 45432	7b. ADDRESS (*City, State, and ZIP Code*) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (*If applicable*) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F41689-84-D-0002
8c. ADDRESS (*City, State, and ZIP Code*) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 62730F; PROJECT NO.: 7719; TASK NO: 18; WORK UNIT ACCESSION NO: 40

11. TITLE (*Include Security Classification*)
Appropriateness Measurement

12. PERSONAL AUTHOR(S)
Drasgow, F.; Levine, M.V.; McLaughlin, M.E.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (*Year, Month, Day*)	15. PAGE COUNT
Final	FROM Jul 84 TO Dec 85	August 1987	138

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	appropriateness indices aptitude test Armed Services Vocational Aptitude Battery
05	08		
05	09		

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

Cheating to raise scores (e.g., to qualify for some desired job or training) and deliberately missing test items to lower scores (e.g., to receive an exemption from military service in a period of general mobilization) are both plausible threats to the integrity of multiple-choice tests. The goal of Appropriateness Measurement is to identify such aberrant test responding; the usual practive is the application of a mathematical procedure to an examinee's item responses which assigns a number (index) related to the probability of aberrant responding. Eleven appropriateness indices were investigated. Three Item Response Theory indices (Drasgow, Levine, and William's l-naught and Tatsuoka's extended caution indices T2 and T4) were effective in detecting aberrant response patterns across a fairly wide range of conditions for a long (85-item) unidimensional test. Their effectiveness was much reduced on a short (30-item) unidimensional test. Methods were developed for combining information across several short unidimensional tests such as are typically found in aptitude batteries, and detection rates were obtained that were comparable to those for the long test. It is concluded that appropriateness indices based on Item Response Theory can be used effectively in operational test programs.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Office	22b. TELEPHONE (*Include Area Code*) (512) 536-3877 22c. OFFICE SYMBOL AFHRL/TSR

DD Form 1473, JUN 86 *Previous editions are obsolete*

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

AFHRL Technical Paper 87-6

August 1987

APPROPRIATENESS MEASUREMENT

Fritz Drasgow
Michael V. Levine
Mary E. McLaughlin

Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432

James A. Earles

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

John R. Welsh, Major, USAF
Chief, Enlisted Selection and Classification Function

Submitted for publication by

Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch
Manpower and Personnel Division

SUMMARY

The military services have a vital concern in assuring that aptitude test scores are appropriate measures of examinees' true abilities. Substantial bonuses have been paid to examinees with sufficiently high scores as enticement to enlist into selected occupations. Under mobilization, exemption from service will be given to examinees with unacceptably low scores. Therefore, cheating to improve scores and deliberately picking incorrect answers to lower scores are both plausible threats to the integrity of enlistment testing. The goal of Appropriateness Measurement is to develop ways to analyze examinees' responses to multiple-choice tests so as to identify such inappropriate test responding.

This effort evaluates 11 practical appropriateness indices. Three, which are based on modern test theory (Item Response Theory), were found to effectively detect aberrant response patterns across a fairly wide range of conditions. This success was obtained when the test had many items but was substantially lessened for military selection test lengths. However, methods developed for combining information on aberrant responding across several different tests resulted in an effectiveness comparable to that found with the longer tests.

The results strongly suggest that appropriateness indices can be used effectively in operational settings. Further research is suggested on a class of indices called "optimal" which hold the promise of even better identification of aberrant responding than those indices already identified.

PREFACE

This effort was accomplished under Project 7719, "Development and Validation of Selection Methodologies." It represents the continuing effort of the Air Force Human Resources Laboratory to fulfill its research and development (R&D) responsibilities through development and application of state-of-the-art methodologies for the continued improvement of the Armed Services Vocational Aptitude Battery (ASVAB).

We wish to thank Bruce Williams and Gregory L. Candell for their help in conducting the research described in Chapter III. They will be coauthors of the paper summarizing this research when it is submitted for journal publication.

TABLE OF CONTENTS

		Page
I.	INTRODUCTION	1
II.	DETECTING INAPPROPRIATE TEST SCORES ON A LONG UNIDIMENSIONAL TEST WITH OPTIMAL AND PRACTICAL APPROPRIATENESS INDICES	3
	Introduction	3
	Appropriateness Indices	4
	Optimal Indices	4
	Standardized ℓ_o	6
	Fit Statistics	7
	Likelihood Function Curvature Statistics	7
	Item-Option Variance	9
	Caution Indices	9
	Standardization	11
	Problem	11
	ROC Curves	11
	Method	12
	Results	13
	Power	13
	Problem	13
	Method	13
	Results	19
	Discussion	28
III.	POLYCHOTOMOUS ANALYSIS OF THE ARITHMETIC REASONING TEST: AN APPLICATION OF MULTILINEAR FORMULA SCORE THEORY	30
	Introduction	30
	Review of Multilinear Formula Score Theory	31
	Estimation and Information	35
	Appropriateness Measurement for the AR Subtest	39
	Purpose	39
	Overview	39
	Appropriateness Indices	41
	Method	42
	Results	43
	Discussion	52

IV. MULTI-TEST EXTENSIONS OF PRACTICAL AND OPTIMAL APPROPRIATENESS INDICES 53

Introduction 53

Multi-Test Extensions of Practical Appropriateness Indices 54

Approximations to Optimal Appropriateness Indices 57

Study One: Simulated ASVAB Data 60

Study Two: Actual ASVAB Data 72

Discussion 76

V. DISCUSSION 85

REFERENCES 89

APPENDIX A: GOODNESS OF FIT OF AR COCCS ESTIMATED FROM A SAMPLE OF N=2,891 AND EVALUATED USING THE ENTIRE SAMPLE OF N=11,914 . . 93

APPENDIX B: ESTIMATED COCCS, SIMULATION COCCS, AND EMPIRICAL PROPORTIONS FROM ESTIMATION SAMPLE 109

APPENDIX C: MULTITEST EXTENTIONS OF OPTIMAL INDICES 125

LIST OF TABLES

Table — Page

1 Ability Distributions Used to Generate Aberrant Samples 18

2 Selected ROC Curve Points for the 15% Spuriously High Treatment, Aberrant Response Patterns Generated from 0-9% Ability Range 20

3 Selected ROC Curve Points for Aberrant Response Patterns Generated from the 0-9% Ability Range 22

4 Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 10-30% Ability Range . . . 23

5 Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 31-48% Ability Range . . . 24

6 Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 49-64% Ability Range . . . 25

7 Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 65-92% Ability Range . . . 26

8	Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 93-100% Ability Range . .	27
9	Means and Standard Deviations of Empirical Distributions of z_3, F2, and T4	29
10	Selected ROC Points for Spuriously High Response Patterns Generated from the 0-9% Ability Range	44
11	Selected ROC Points for Spuriously High Response Patterns Generated from the 10-30% Ability Range . . .	45
12	Selected ROC Points for Spuriously High Response Patterns Generated from the 31-48% Ability Range . . .	46
13	Selected ROC Points for Spuriously High Response Patterns Generated from the 49-64% Ability Range . . .	47
14	Selected ROC Points for Spuriously Low Response Patterns Generated from the 31-48% Ability Range . . .	48
15	Selected ROC Points for Spuriously Low Response Patterns Generated from the 49-64% Ability Range . . .	49
16	Selected ROC Points for Spuriously Low Response Patterns Generated from the 65-92% Ability Range . . .	50
17	Selected ROC Points for Spuriously Low Response Patterns Generated from the 93-100% Ability Range . .	51
18	Selected ROC Points for Spuriously High Response Patterns Generated from the 00-09% Ability Range . . .	64
19	Selected ROC Points for Spuriously High Response Patterns Generated from the 10-30% Ability Range . . .	65
20	Selected ROC Points for Spuriously High Response Patterns Generated from the 31-48% Ability Range . . .	66
21	Selected ROC Points for Spuriously High Response Patterns Generated from the 49-64% Ability Range . . .	67
22	Selected ROC Points for Spuriously Low Response Patterns Generated from the 31-48% Ability Range . . .	68
23	Selected ROC Points for Spuriously Low Response Patterns Generated from the 49-64% Ability Range . . .	69
24	Selected ROC Points for Spuriously Low Response Patterns Generated from the 65-92% Ability Range . . .	70
25	Selected ROC Points for Spuriously Low Response Patterns Generated from the 93-100% Ability Range . .	71

26 Selected ROC Points for Spuriously High Response Patterns Created from NORC Examinees in the 00-09% Ability Range 76

27 Selected ROC Points for Spuriously High Response Patterns Created from NORC Examinees in the 10-30% Ability Range 77

28 Selected ROC Points for Spuriously High Response Patterns Created from NORC Examinees in the 31-48% Ability Range 78

29 Selected ROC Points for Spuriously High Response Patterns Created from NORC Examinees in the 49-64% Ability Range 79

30 Selected ROC Points for Spuriously Low Response Patterns Created from NORC Examinees in the 31-48% Ability Range 80

31 Selected ROC Points for Spuriously Low Response Patterns Created from NORC Examinees in the 49-64% Ability Range 81

32 Selected ROC Points for Spuriously Low Response Patterns Created from NORC Examinees in the 65-92% Ability Range 82

33 Selected ROC Points for Spuriously Low Response Patterns Created from NORC Examinees in the 93-100% Ability Range 83

LIST OF ILLUSTRATIONS

Figure Page

1 ROC curves obtained from 200 normal low θ response vectors and 200 normal average θ response vectors . . 14

2 ROC curves obtained from 200 normal average θ response vectors and 200 normal high θ response vectors 15

3 ROC curves obtained from 200 normal low θ response vectors and 200 normal high θ response vectors 16

4 Ability density for National Opinion Research Center sample on the Arithmetic Reasoning subtest 36

5 Information functions for dichotomous and polychotomous scorings of the Arithmetic Reasoning subtest 40

6 Likelihood ratios evaluated by Simpson's rule and the quadratic approximation for simulated normal response patterns 59

7 Standardizations of practical appropriateness indices . 75

8 Detection rates of z_3, T2, and T4 expressed as proportions of the rate of the optimal index at a 1% false alarm rate 86

I. INTRODUCTION AND OVERVIEW

Some examinees' scores on a multiple-choice test may fail to provide valid measures of the trait measured by the test. Examinees can obtain spuriously high scores because they copy answers from more talented neighbors or because they have been given the answers to some questions. Examinees can obtain spuriously low scores due to alignment errors (answering, say, the tenth item in the space provided for the ninth item, answering the eleventh item in the space provided for the tenth item, etc.), language difficulties, atypical educations, and unusually creative interpretations of normally easy items.

Detecting inappropriate test scores is very important in military testing. For example, substantial recruitment bonuses may be erroneously paid to low ability examinees who obtain spuriously high test scores. Many of these individuals are likely to fail to complete military technical training schools; this leads to high attrition costs. Even when such individuals are able to complete training, they are likely to exhibit low on-the-job performances.

Spuriously low scores can also cause serious difficulties in military testing. Spuriously low scores can lead to difficulties in filling important manpower needs because truly able individuals will be inappropriately disqualified. This problem is likely to be exacerbated in the future as the birthrates of many demographic groups decline.

The goal of Appropriateness Measurement is to identify inappropriate test scores. In recent years, several methods for identifying these test scores have been devised. In all approaches, response patterns are characterized in a way that permits us to assess quantitatively the degree to which an observed response vector is atypical. This quantitative measure is then used to classify response patterns into appropriate (i.e., normal) and inappropriate (i.e., aberrant) categories.

In a series of studies, it has been found that simulated spuriously high response patterns and simulated spuriously low response patterns can be detected by appropriateness measurement. High detection rates have been obtained despite model misspecification, errors in item characteristic curve parameter estimates, and the inclusion of inappropriate response patterns in the test norming sample (Levine & Drasgow, 1982). Very high detection rates have been obtained when response patterns of low ability examinees have been modified to simulate cheating and when response patterns of high ability examinees have been modified to simulate spuriously low responding (Drasgow, Levine, & Williams, 1985).

Among the many methods that have been proposed, which is best for detecting inappropriate test scores on the short unidimensional power subtests from the Armed Services Vocational Aptitude Battery (ASVAB)? Also, is there some clearly superior method that has not yet been proposed? Previous

research on Appropriateness Measurement has generally focused on long unidimensional tests such as the Scholastic Aptitude Test-Verbal section (SAT-V) and the Graduate Record Examination-Verbal section (GRE-V). The research described in this paper was designed to determine which of these indices is best for ASVAB subtests (in particular, the portion known as the Armed Forces Qualification Test or AFQT) and, as described below, to decide if the best method currently available could be significantly improved.

The difficult problem of evaluating the effectiveness of an appropriateness index was recently solved to a large extent by Levine and Drasgow (1984; 1987). They developed statistical theory and numerical methods that enabled them to compute optimal appropriateness indices for given forms of aberrance. These indices are optimal in the sense that no other statistics computed from an examinee's item responses can achieve higher rates of detection (at each error level) of given forms of aberrance. Thus, the absolute effectiveness of any practical, easy-to-compute appropriateness index previously suggested in the literature can be determined by comparing it to an optimal index.

Many appropriateness indices were evaluated in the present effort. The best practical appropriateness indices based on Item Response Theory (IRT) were found to be far superior to non-IRT alternatives, such as the standardized residual from a multiple regression equation. In some cases, the best practical indices had detection rates that were nearly as high as the detection rates of optimal appropriateness indices. In other situations, optimal indices provided far higher detection rates.

At present, optimal indices show promise for use in operational settings. With further development, optimal indices could be used to provide powerful detection of specific forms of aberrance that are difficult to detect using even the best practical indices. For example, suppose a test score falls into AFQT Category 3A. Does the examinee truly belong to this ability category? Or is the examinee actually an AFQT Category 3B examinee who was unethically given the answers to a moderate number of items? An optimal index can be formulated to test such hypotheses.

In the first study described in this report, 11 practical appropriateness indices were evaluated and compared to optimal indices. Simulated SAT-V data were used in the first study because many of the practical indices were originally proposed in the context of a long unidimensional test. Optimal indices were found to provide very high rates of detection of inappropriate response patterns. The best practical indices were nearly optimal in some conditions but fell short of optimal in other conditions.

In the second study conducted for this effort, the effectiveness of each of the practical and optimal indices on a short unidimensional test was evaluated using simulated ASVAB Arithmetic Reasoning (AR) subtest data. Rates of detection of aberrant response patterns were found to be substantially reduced for the short AR subtest in relation to the long SAT-V test.

Methods were then developed for combining information about aberrance across several short unidimensional tests. Simulated and actual ASVAB data for the AR, Word Knowledge (WK), and Paragraph Comprehension (PC) subtests were used to evaluate the multi-test appropriateness indices. By increasing

the number of items from 30 on the AR test to 80 on the combined AR, WK, and PC subtests, we obtained detection rates that were comparable to the 85-item SAT-V.

The following chapters describe the present research and development (R&D) effort, provide concluding remarks, and suggest directions for future R&D. The results strongly suggest that appropriateness indices based on IRT can be used effectively in operational settings. Further significant gains in detection rates are expected if optimal indices are developed for use in operational settings.

II. DETECTING INAPPROPRIATE TEST SCORES ON A LONG UNIDIMENSIONAL TEST WITH OPTIMAL AND PRACTICAL APPROPRIATENESS INDICES

Introduction

It is relatively easy to propose new appropriateness indices. Unfortunately, evaluations of the relative merits of the various indices have been very difficult in previous research. Cliff's (1979, p. 388) description of a related problem cogently summarized the difficulty in evaluating indices: "Now the trouble is that the formulas multiply not just like rabbits, or even guppies, but rather like amoebae: by both fusion and conjugation, and there seemed to be no general principle to use in selecting from among them." Harnisch and Tatsuoka (1983), for example, correlated 14 different indices in order to see which pairs were more and less related, but this approach has limited value in determining which index is best. Furthermore, this approach does not determine which indices, if any, are good enough for operational use.

In the past, two criteria have been used to evaluate appropriateness indices: standardization and relative power. Standardization, introduced by Drasgow, Levine, and Williams (1985), refers to the extent to which the conditional distributions (given particular values of the latent trait) of an index are invariant across levels of the latent trait. There is little confounding between ability and measured appropriateness for a well-standardized index. Well standardized indices have two attractive features. First, high rates of detection of aberrant response patterns by well-standardized indices cannot be due merely to differences in ability distributions or number-right distributions across normal and aberrant samples. In contrast, high detection rates obtained by poorly standardized indices may be due largely to differences in ability distributions. This point is illustrated in a later section of this chapter. Second, a well-standardized index is easy to use in practice because index scores for individuals with different standings on the latent trait can be compared directly. In contrast, scores on poorly standardized indices can be interpreted only in relation to their conditional distributions; consequently, a single cutting score for classification into aberrant and appropriate groups is not possible. Furthermore, it is sometimes very difficult and time-consuming to obtain the conditional distributions of an appropriateness index. In such cases, the practical usefulness of the index is limited.

Relative power, the second criterion used to evaluate appropriateness indices, refers to the ability of a particular index to correctly classify aberrant response patterns as aberrant, compared to the classification rate of another index. If some well-standardized index has acceptable power, then it can be used in operational settings. Unfortunately, no unequivocal conclusions about the detectability of some form of aberrance are possible if none of the indices under consideration has adequate power. We do not know whether or not there exists some other index, not included in the experimental study, that has acceptable power. In addition, even if an index were found to have adequate power for operational use, we do not know whether or not there is an index, as yet undiscovered, that is substantially superior to all known indices.

It is now possible to determine the detectability of a specified form of aberrance by the methods devised by Levine and Drasgow (1984; 1987). They introduced a general method for ascertaining the maximum power that can be achieved by any index. Chapters 2, 3, and 4 contain the first major applications of the method.

By means of a new numerical algorithm, Levine and Drasgow (1984; 1987) were able to apply the Neyman-Pearson Lemma to specify an appropriateness index that is optimal in the sense that no other index computed from the item responses can achieve a higher detection rate (at each error rate) of the given form of aberrance.

As a result of their research, it is now possible to determine the absolute effectiveness of an index for detecting a particular type of aberrance on a given test. The absolute effectiveness of an index is determined by comparing its detection rate with the detection rate of the corresponding optimal index. In the first study conducted for this effort, 11 different appropriateness indices were evaluated for their abilities to detect spuriously high and low response patterns on a long unidimensional power test: namely, the SAT-V.

The appropriateness indices examined in the first study and some computational notes are presented in the next section. The extent to which each index is standardized and the power of each index for detecting several forms of aberrance are then examined. Some remarks concerning the results are provided in the final section of this chapter.

Appropriateness Indices

Optimal Indices

Suppose we wish to test a simple null hypothesis against a simple alternative hypothesis. If the probability of a Type I error is α, then the most powerful test is the test that minimizes the probability of a Type II error among the set of tests with the given Type I error rate. The Neyman-Pearson Lemma states that maximum power is achieved by a likelihood ratio test. More specifically, let $L_N(\mathbf{x})$ and $L_A(\mathbf{x})$ denote the likelihoods of the data $\mathbf{x}$ under the null and alternative hypotheses, respectively. Then the Neyman-Pearson Lemma states that of all tests with a Type I error rate of α,

none is more powerful than a test obtained from the likelihood ratio $\underline{L}_A(\mathbf{x})/\underline{L}_N(\mathbf{x})$.

The Neyman-Pearson Lemma can be applied in the context of Appropriateness Measurement to construct most powerful tests and, consequently, optimal appropriateness indices. To see how it is used, suppose that local independence holds, $\mathbf{u} = (\underline{u}_1, \ldots, \underline{u}_n)$, and $\underline{P}_i(\underline{u}_i|\theta)$ is the probability of response $\underline{u}_i$ to item $\underline{i}$ by an examinee of ability θ under the null hypothesis that the response pattern is appropriate (normal). Then the likelihood of a response vector $\mathbf{u}$ by an examinee of ability θ is

$$\underline{P}_{\text{Normal}}(\mathbf{u}|\theta) = \prod_{i=1}^{n} \underline{P}_i(\underline{u}_i|\theta).$$

If the ability density is $\underline{f}(\theta)$, then using elementary probability

$$\underline{P}_{\text{Normal}}(\mathbf{u}) = \int P_{\text{Normal}}(\mathbf{u}|\theta)\ \underline{f}(\theta)\ d\theta.$$

To apply the Neyman-Pearson Lemma, it is necessary to compute $\underline{P}_{\text{Aberrant}}(\mathbf{u})$. This quantity can be obtained by carrying the conditioning-integrating argument one step further. For concreteness, suppose that the type of aberrance under consideration consists of $\underline{m}$ randomly selected items being modified by the spuriously low treatment. Let $\underline{S}_k$ denote a set indicating the $\underline{k}$th way of selecting $\underline{m}$ of $\underline{n}$ items (of the $\binom{n}{m}$ ways possible), let $\underline{P}_{\text{Aberrant}}(\mathbf{u}|\theta,\underline{S}_k)$ denote the likelihood of response pattern $\mathbf{u}$ for an examinee with ability θ when the items in $\underline{S}_k$ are subjected to the spuriously low treatment, and let $\underline{P}(\underline{S}_k)$ denote the probability of $\underline{S}_k$ (i.e., $\underline{P}(\underline{S}_k) = 1/\binom{n}{m}$). Then

$$\underline{P}_{\text{Aberrant}}(\mathbf{u}|\theta) = \sum_k \underline{P}_{\text{Aberrant}}(\mathbf{u}|\theta,\underline{S}_k)\ \underline{P}(\underline{S}_k)$$

so that

$$\underline{P}_{\text{Aberrant}}(\mathbf{u}) = \int \left[\sum \underline{P}_{\text{Aberrant}}(\mathbf{u}|\theta,\underline{S}_k)\ \underline{P}(\underline{S}_k)\right] \underline{f}(\theta)\ d\theta. \qquad (1)$$

By taking advantage of the symmetry in the $\underline{P}_{\text{Aberrant}}(\mathbf{u}|\theta,\underline{S}_k)$, Levine and Drasgow (1984) obtained an efficient numerical algorithm for computing $\underline{P}_{\text{Aberrant}}(\mathbf{u}|\theta)$. Using a numerical quadrature formula, the right-hand side of Equation 1 can be accurately evaluated with an acceptable amount of computation. Details about these calculations and a theoretical treatment of the general problem are provided by Levine and Drasgow (1984; 1987).

Thus, it is possible to compute the likelihood ratio

$$LR = \underline{P}_{Aberrant}(u)/\underline{P}_{Normal}(u) \qquad (2)$$

and test the simple null hypothesis that a response pattern is normal against the simple alternative hypothesis that the response pattern is aberrant. Due to the Neyman-Pearson Lemma, the likelihood ratio statistic provides a most powerful test; consequently, when it is used as an appropriateness index, the likelihood ratio statistic is as powerful as any index that can be computed from the item responses.

Standardized ℓ_o

Let z_3 denote the standardized ℓ_o index (Drasgow et al., 1985). It may be computed by the formula

$$z_3 = \frac{\ell_o - M(\hat{\theta})}{[S(\hat{\theta})]^{1/2}} . \qquad (3)$$

In this formula, ℓ_o is the logarithm of the three-parameter logistic likelihood function evaluated at the maximum likelihood estimate $\hat{\theta}$ of θ:

$$\ell_o = \sum_{i=1}^{n} [\underline{u}_i \log \underline{P}_i(\hat{\theta}) + (1-\underline{u}_i)\log \underline{Q}_i(\hat{\theta})],$$

where $\underline{u}_i$ is the dichotomously scored (1=correct, 0=incorrect) item response for item $\underline{i}$, $\underline{i}$ = 1, 2, ..., $\underline{n}$; $\underline{Q}_i(\theta) = 1 - \underline{P}_i(\theta)$;

$$\underline{P}_i(\theta) = \hat{\underline{c}}_i + \frac{1 - \hat{\underline{c}}_i}{1 + \exp[-\underline{D}\hat{\underline{a}}_i(\theta-\hat{\underline{b}}_i)]} ; \qquad (4)$$

$\underline{D}$ = 1.702; and $\hat{\underline{a}}_i$, $\hat{\underline{b}}_i$, and $\hat{\underline{c}}_i$ are item parameter estimates.

The conditional expectation of ℓ_o, given $\theta = \hat{\theta}$, is

$$M(\hat{\theta}) = \sum_{i=1}^{n} [\underline{P}_i(\hat{\theta}) \log \underline{P}_i(\hat{\theta}) + \underline{Q}_i(\hat{\theta}) \log \underline{Q}(\hat{\theta}_i)] \qquad (5)$$

and its conditional variance is

$$S(\hat{\theta}) = \sum_{i=1}^{n} [\underline{P}_i(\hat{\theta})\underline{Q}_i(\hat{\theta})[\log(\underline{P}_i(\hat{\theta})/\underline{Q}_i(\hat{\theta}))]^2]. \qquad (6)$$

Justifications of these formulas can be found in Drasgow et al. (1985).

Fit Statistics

Two fit statistics for the three-parameter logistic model were suggested by Rudner (1983) as generalizations of the Rasch model fit statistics used by Wright (1977) and his colleagues. The first is the mean squared standardized residual

$$F1 = \frac{1}{n} \sum_{i=1}^{n} [u_i - P_i(\hat{\theta})]^2 / [P_i(\hat{\theta})Q_i(\hat{\theta})]. \quad (7)$$

The other fit statistic is

$$F2 = \sum_{i=1}^{n} [u_i - P_i(\hat{\theta})]^2 / \sum_{i=1}^{n} P_i(\hat{\theta})Q_i(\hat{\theta}), \quad (8)$$

which Rudner found to be quite effective in some cases (see Rudner, 1983, p. 214 and p. 216, where Rudner uses W3 to denote an expression proportional to Equation 8).

Likelihood Function Curvature Statistics

Four indices that provide measures of the "flatness" of the likelihood function were evaluated. These indices are motivated by the notion that inappropriate responses will flatten the likelihood function near its maximum because no single value of θ will allow the item response model to provide a good fit to the response vector. Therefore, the likelihood function will not have a sharp maximum; instead, it will be relatively flat.

Normalized Jackknife. The first measure of the curvature of the likelihood function is the normalized jackknife variance estimate. In order to compute this index, let $\hat{\theta}$ denote the three-parameter logistic maximum likelihood estimate of ability based on all n test items and let $\hat{\theta}_{(j)}$ denote the estimate based on the $n - 1$ items remaining when item j is excluded. The pseudo-values (see, for example, Mosteller & Tukey, 1968) are

$$\hat{\theta}_j^* = n\hat{\theta} - (n-1)\hat{\theta}_{(j)}^*, \quad j = 1, 2, \ldots, n.$$

The jackknife estimate of θ is then

$$\hat{\theta}^* = \frac{1}{n} \sum_{j=1}^{n} \hat{\theta}_j^*$$

and the jackknife estimate of its variance is

$$\mathrm{Var}(\hat{\theta}^*) = \frac{\sum(\hat{\theta}_j^*)^2 - \frac{1}{n}(\sum \hat{\theta}_j^*)^2}{n(n-1)}.$$

The jackknife variance estimate is not a standardized appropriateness index; there is more Fisher information about θ in some ability ranges than in others, and so $\mathrm{Var}(\hat{\theta}^*)$ is expected to depend upon θ. Lord's (1980) formula for the information of the three-parameter logistic maximum likelihood estimate of θ,

$$\underline{I}(\theta) = \sum_{i=1}^{n} \frac{[\underline{P}_i(\theta)']^2}{\underline{P}_i(\theta)\underline{Q}_i(\theta)}, \tag{9}$$

can be used to reduce this problem. Since the reciprocal of $\underline{I}(\theta)$ is the asymptotic variance of $\hat{\theta}$, the jackknife estimate of variance can be approximately normalized by evaluating the information function at $\hat{\theta}$ and computing

$$JK = \mathrm{Var}(\hat{\theta}^*)\underline{I}(\hat{\theta}). \tag{10}$$

It is possible to arrange the calculations for computing JK very efficiently. We found that one Newton-Raphson iteration was adequate to move from $\hat{\theta}$ to $\hat{\theta}_{(j)}$. Then, since the first and second derivatives of the log likelihood functions for the whole test are sums over $\underline{n}$ items, the first and second derivatives of the log likelihood functions for the $\underline{n}$-1 item test can be obtained by single subtractions of already computed quantities. Consequently, all the pseudo-values, $\hat{\theta}^*$, and JK can be obtained with fewer arithmetic calculations than are required in a single Newton-Raphson iteration in the calculation of $\hat{\theta}$.

Convergence of $\hat{\theta}$. A possible consequence of a relatively flat likelihood function for aberrant response patterns is that the number of iterations required to obtain $\hat{\theta}$ may be increased. The number NI of Newton-Raphson iterations required to obtain $\hat{\theta}$ can therefore be used as an appropriateness index.

Expected versus Observed Likelihood Function Curvatures. This index (O/E) is also motivated by an hypothesis about the likelihood function's flatness. If the likelihood function is flatter for aberrant response patterns than for normal response patterns, then we would expect that the observed information, defined as minus the second derivative of the log likelihood function at $\hat{\theta}$ given the response vector **u** (see Efron & Hinkley, 1978, p. 457), would be less than the information $\underline{I}(\hat{\theta})$ given in Equation 9, which (given $\hat{\theta}$) does not depend upon **u**. Thus, the sixth index is the ratio of the observed and expected information

$$O/E = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\hat{\theta}} \Big/ \underline{I}(\hat{\theta}), \tag{11}$$

where ℓ is the log likelihood

$$\ell = \sum_{i=1}^{n} [u_i \log P_i(\theta) + (1-u_i)\log Q_i(\theta)]. \qquad (12)$$

Bayes Posterior Variance. Another statistic closely related to the O/E index is the posterior variance B of the Bayes estimate of θ. It is expected to be relatively large for aberrant response vectors and relatively small for normal response vectors. Thus, it should serve to distinguish between normal and aberrant response patterns.

Item-Option Variance

Suppose that we consider the subset of N_{ik} examinees in the test norming sample who selected option k to item i. It is easy to compute the mean number-right score $\bar{X}_{ik}$ for these examinees. In this way, we can identify options to item i that are typically selected by high ability examinees (e.g., the correct option) and options that are typically selected by lower ability examinees. For spuriously high and low response patterns, we would expect to observe inconsistency in $\bar{X}_{ik}$; sometimes options with low $\bar{X}_{ik}$ are selected and sometimes options with high $\bar{X}_{ik}$ are selected. For this reason, we evaluated the item-option variance

$$IOV = Var(\bar{X}_{ik}) \qquad (13)$$

as a measure of appropriateness.

Caution Indices

Sato's Caution Index. Three "caution indices" were also be examined. The first is Sato's (1975) caution index S (see also Tatsuoka & Linn, 1983, but replace $y_{\bullet j}$ with $P_{\bullet j}$ for a simpler version of their Equation 1). S is easy to compute and is widely used in Japan. To compute S, suppose that the n test items are ordered from easiest to hardest on the basis of proportion right $\hat{p}_i$ in the test norming sample. Let

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i$$

be the mean proportion correct and suppose that an examinee answers k items correctly. If $\hat{\mathbf{p}}$ is a vector containing the $\hat{p}_i$ and $\mathbf{g}$ is a perfect Guttman response pattern with 1s as its first k elements and 0s for the next $n-k$ elements, then

$$S = 1 - \frac{\mathrm{Cov}(u, \hat{p})}{\mathrm{Cov}(g, \hat{p})}$$

$$= 1 - \frac{\sum_{i=1}^{n} u_i(\hat{p}_i - \bar{p})}{\sum_{i=1}^{k} (\hat{p}_i - \bar{p})} . \qquad (14)$$

Note that the summation in the denominator of the last expression is from 1 to k (i.e., over the k items with the smallest $\hat{p}_i$ values), not 1 to n.

Tatsuoka's Standardized Extended Caution Indices. Two indices that are related to Sato's caution index are the second and fourth standardized extended caution indices T2 and T4 presented by Tatsuoka (1984, p. 104). These two indices (of the four studied by Tatsuoka) were included because Harnisch and Tatsuoka (1983) found that these indices were not related (linearly or curvilinearly) to true score and, therefore, $\hat{\theta}$.

T2 and T4 can be computed relatively easily. Let $\hat{\theta}_j$ denote the three-parameter logistic maximum likelihood estimate of ability for the jth person in the test norming sample of N examinees, and let $P_{ij}(\hat{\theta}_j)$ be the probability of a correct response to item i by this person computed from Equation 4. Then define

$$G_i = \frac{1}{N} \sum_{j=1}^{N} P_{ij}(\hat{\theta}_j)$$

and

$$\bar{G} = \frac{1}{n} \sum_{i=1}^{n} G_i .$$

To compute T2 and T4 for an examinee in the normal sample or an aberrant sample, let

$$\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i(\hat{\theta}) .$$

Then

$$T2 = \frac{\sum [(\underline{P}_i(\hat{\theta}) - \underline{u}_i)(\underline{G}_i - \bar{\underline{G}})}{[\sum \underline{P}_i(\hat{\theta})\underline{Q}_i(\hat{\theta})(\underline{G}_i - \bar{\underline{G}})^2]^{1/2}} \quad (15)$$

and

$$T4 = \frac{\sum [(\underline{P}_i(\hat{\theta}) - \underline{u}_i)(\underline{P}_i(\hat{\theta}) - \bar{\underline{P}})}{[\sum \underline{P}_i(\hat{\theta})\underline{Q}_i(\hat{\theta})(\underline{P}_i(\hat{\theta}) - \bar{\underline{P}})^2]^{1/2}} . \quad (16)$$

It should be noted that Equations 14, 15, and 16 are generalizations of the original caution indices to the situation where item parameters are estimated in a test norming sample.

Standardization

Problem

Measured appropriateness can be confounded with ability. Drasgow et al. (1985, p. 74), for example, provide an example of a strong, nearly linear relation between estimated ability and an unstandardized index. A score of, say, -50 on this index at one ability level indicates a good fit of the model to a response vector, but the same index score at other ability levels indicates a very poor fit. Consequently, an observed difference between the distributions of index scores for normal and aberrant response vectors is not unequivocal evidence of index effectiveness. Instead, it may simply reflect differences in ability or number-right distributions. This problem does not occur if an appropriateness index is well standardized; that is, if the conditional distributions (given θ) of the index are (approximately) equal across possible values of θ for normal examinees.

In practical applications of Appropriateness Measurement, it would be convenient if a single cutting score could be used to classify response patterns as aberrant or normal. If the conditional distributions of an index are not identical, then the interpretation of a score on a practical appropriateness index must be made vis-a-vis the associated conditional distribution. Consequently, it would not be possible to use a single cutting score to classify response patterns as aberrant nor would it be possible to compare directly index scores of examinees with differing abilities.

We would expect little degradation of the performance of a well-standardized index if the ability distribution were to change abruptly. Such a change would be expected, for example, with the ASVAB examinee population in a period of national mobilization.

ROC Curves

If an index is properly standardized, its distribution will be nearly the same in subpopulations of normal examinees who differ in ability. Hence, the index could not be used to distinguish among groups. A standard, very general method for studying the extent to which some statistic can differentiate between two groups is the Receiver Operating Characteristic (ROC) curve.

Thus, we can study index standardization by using an ROC curve to determine whether the index distinguishes between groups of normal examinees who differ in ability.

An ROC curve is obtained by specifying a cutting score $\underline{t}$ for an index and then computing

$\underline{x}(\underline{t})$ = proportion of group 1 (say, normal, low ability examinees) response patterns with index values less than $\underline{t}$ (assuming that small index values indicate aberrance);

$\underline{y}(t)$ = proportion of group 2 (say, normal, high ability examinees) response patterns with index scores less than $\underline{t}$.

An ROC curve consists of the points $(\underline{x}(\underline{t}),\underline{y}(\underline{t}))$ obtained for various values of $\underline{t}$. The proportion $\underline{x}(\underline{t})$ is called the false alarm rate, and $\underline{y}(\underline{t})$ is called the hit rate. A detailed example of the construction of an ROC curve is given by Hulin, Drasgow, and Parsons (1983, pp. 131-134).

An appropriateness index is well-standardized across two ability levels if the ROC curve lies along the diagonal line $\underline{y} = \underline{x}$.

Method

Polychotomous item responses (five-option multiple-choice items with omitting allowed) were simulated using the histograms constructed by Levine and Drasgow (1983). They used the three-parameter logistic model to estimate the abilities of 49,470 examinees from the 85-item April 1975 administration of the SAT-V. Then the examinees were sorted into 25 groups on the basis of estimated ability. The 4th, 8th, ..., 96th percentiles of the normal (0,1) distribution were used as cutting scores when sorting examinees. Then the proportions of examinees choosing each option (treating skipped and not-reached items as a single response category) were computed for each of the 25 ability groups. Probabilities of option choices were then computed by linear interpolation between category medians (i.e., the 2nd, 6th, ..., 98th percentiles from the normal (0,1) distribution).

Five samples of normal response patterns were generated by first sampling 3,000 numbers (θ's) from the normal (0,1) distribution truncated to the [-2.05, 2.05] interval. (It was necessary to truncate the ability distribution because interpolation below the 2nd percentile or above the 98th percentile was not possible with the histograms.) Then low [-2.05 to -1.50), moderately low (-.70 to -.55), average (-.05 to .05), moderately high (.55 to .70), and high (1.49 to 2.05] θ samples of N = 200 each were formed.

Polychotomous item response vectors were then generated for each θ value. For each item, the associated histogram was used to compute the conditional (given θ) probabilities of the six possible responses (treating skipped and not-reached as the sixth response). A number was sampled from the uniform distribution on the unit interval, and a simulated response was obtained by determining where the random number was located in the cumulative distribution corresponding to the conditional probabilities.

Finally, each of the 11 practical appropriateness indices was computed for each response vector in each sample. Then ROC curves were computed for each of the $\binom{5}{2} = 10$ possible pairs of samples and each of the 11 appropriateness indices.

Results

Figures 1 through 3 present the results for the low-average, average-high, and low-high comparisons. The results for the other seven comparisons were consistent with the trends seen in these three figures; consequently, they will not be presented. Furthermore, only the lower left quarter of the ROC curve is plotted because it is unlikely that anyone would set a cutting score that yielded a false alarm rate of more than 50%.

In Figure 1, it is evident that NI, IOV, S, and B are poorly standardized. This result is not surprising because no explicit steps were taken to standardize these indices. The standardizations of the z_z, F1, F2, JK, and O/E indices seem reasonably good across low θ and average θ groups. The standardization of T2 and T4 seem somewhat less adequate, although T2 is well standardized for false alarm rates of less than .20.

The pattern of results in Figure 2 is somewhat different from the results in Figure 1. In both figures, NI, IOV, and B are poorly standardized, and z_z, F2, JK, and O/E are again well standardized. But F1 is much less well standardized in Figure 2. In contrast, the results for S and T4 have improved considerably. The standardization of T2 was better in Figure 1.

Finally, Figure 3 presents the results comparing the low θ normals to the high θ normals. The pattern of results indicates that this comparison is the most severe test of standardization. Note that at low misclassification rates, only z_z, F2, and JK have ROC curves near the diagonal. The standardizations of NI, IOV, F1, B, and S are all poor. T4 seems standardized somewhat better than T2.

Power

Problem

Do any of the well-standardized practical appropriateness indices have adequate power for detecting some form of aberrance? Are any nearly as powerful as the index that is optimal for the given form of aberrance?

Method

Data Sets. A test norming sample of 3,000 response vectors was created by sampling 3,000 numbers (θs) from the normal (0,1) distribution truncated to the [-2.05,2.05] interval. A normal sample of 4,000 response vectors was also generated in this way. Two thousand aberrant response vectors were created in each of 12 conditions. The 12 conditions resulted from varying three factors: the type of aberrance (spuriously high; spuriously low), the severity of aberrance (mild; moderate), and the distribution from which simulated abilities were sampled.

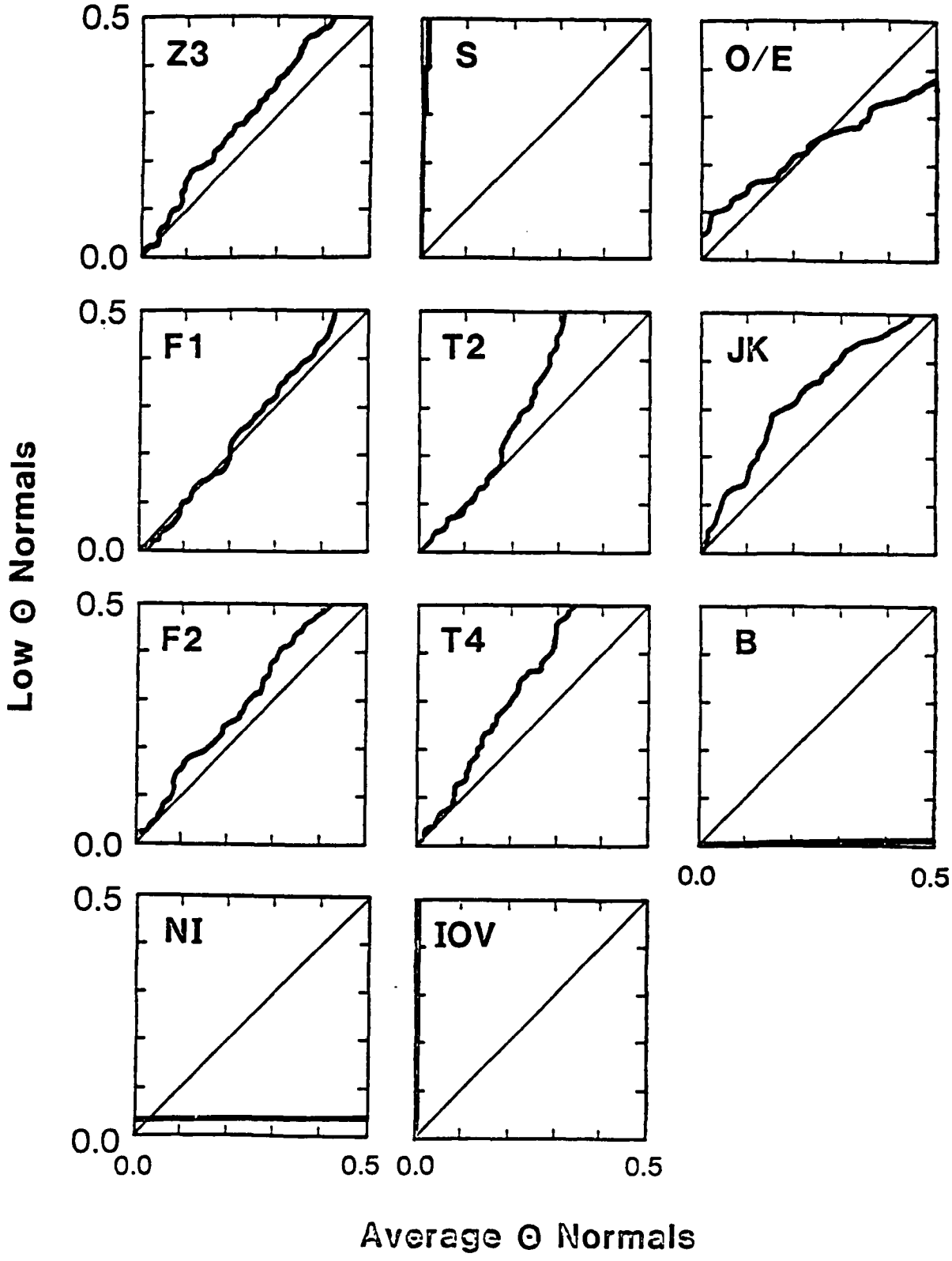


Figure 1. ROC curves obtained from 200 normal low θ response vectors and 200 normal average θ response vectors.

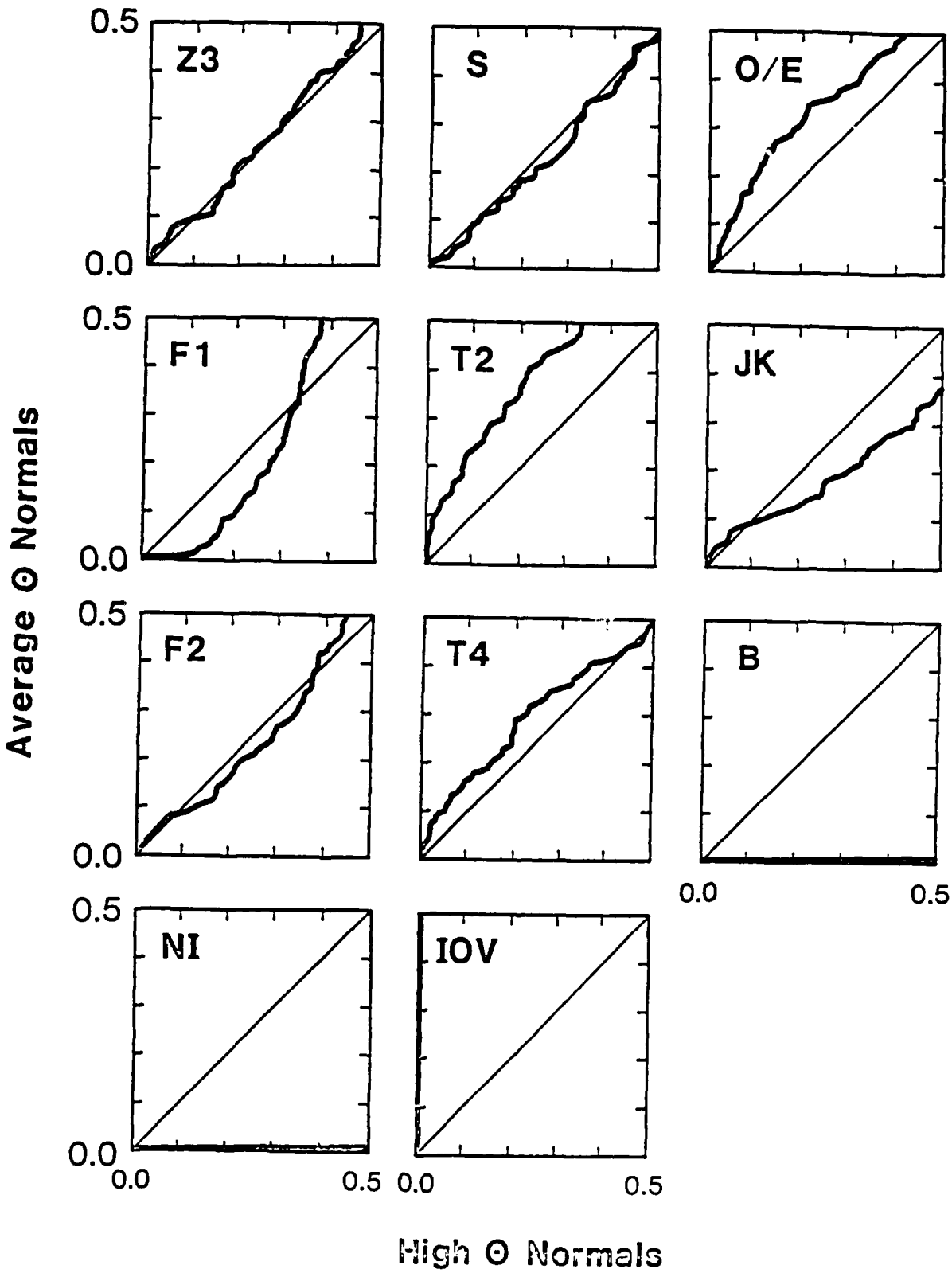


Figure 2. ROC curves obtained from 200 normal average θ response vectors and 200 normal high θ response vectors.

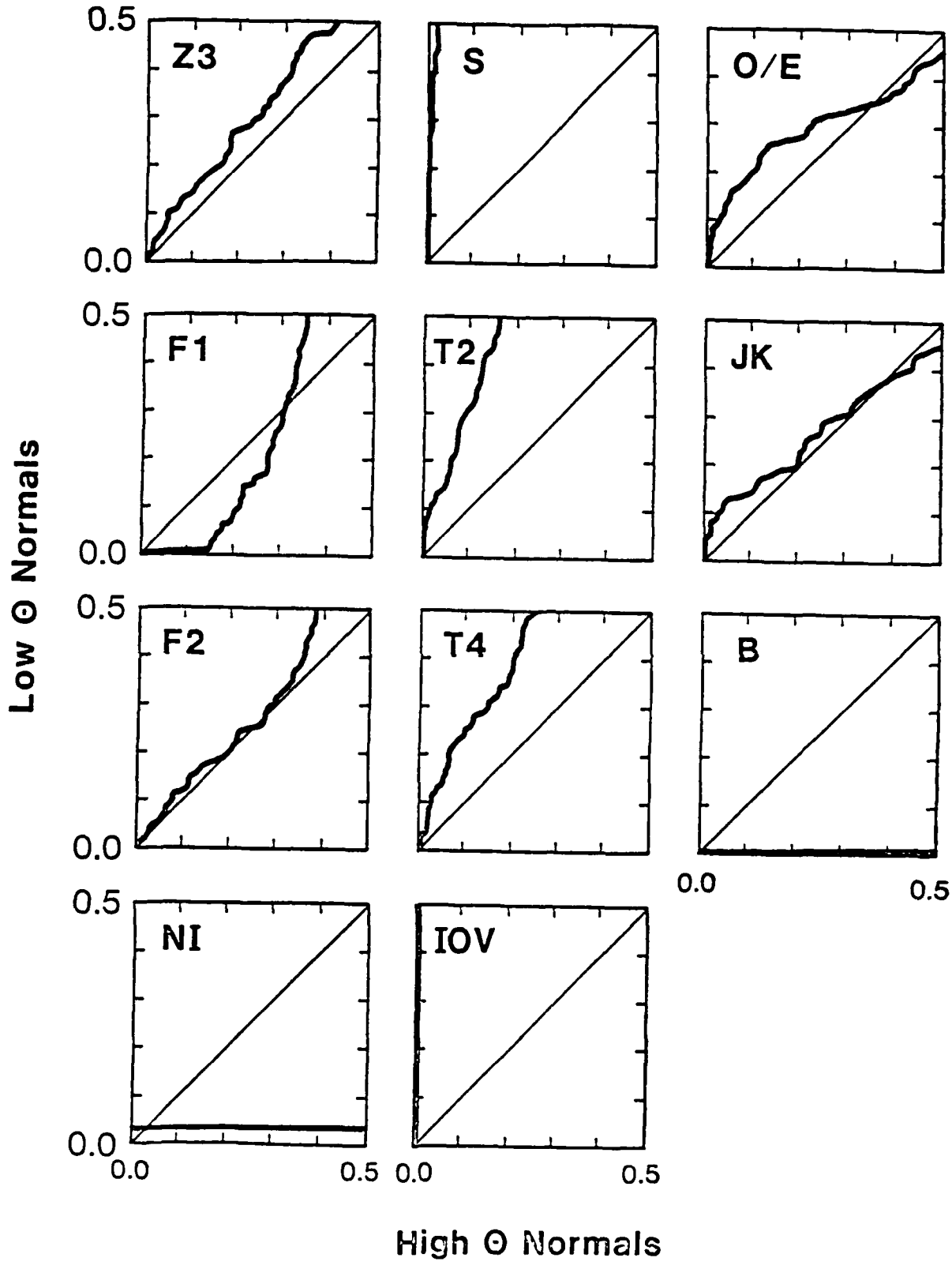


Figure 3. ROC curves obtained from 200 normal low θ response vectors and 200 normal high θ response vectors.

Six of the aberrant samples contained spuriously high response vectors, and the remaining six samples contained spuriously low response vectors. Spuriously high response patterns were created by first generating normal response vectors (polychotomously scored) and then replacing a given percentage k of simulated responses (randomly sampled without replacement) with correct responses. Spuriously low response patterns were also created by first generating normal response vectors. Then a fixed percentage of items were randomly selected without replacement, and the responses to these items were replaced with random responses (i.e., a response was replaced by option A with probability .2, by option B with probability .2, ..., and by option E with probability .2). Mildly aberrant response patterns were generated by using $k = 15\%$. Moderately aberrant response patterns were created using $k = 30\%$.

The third variable manipulated was the ability level of the aberrant sample. Abilities for the spuriously high samples were sampled from three parts of the normal (0,1) distribution truncated to [-2.05,2.05]: very low (0th through 9th percentiles), low (10th through 30th percentiles), and low average (31st through 48th percentiles). In all cases, percentile points were determined after the truncation to [-2.05, 2.05]. These intervals were used because it is more important to detect spuriously high response patterns for low ability examinees than for high ability examinees. Similarly, it is more important to detect spuriously low responses for high ability examinees. Consequently, abilities were sampled from three above-average ability strata for the spuriously low samples: very high (93rd percentile and above), high (65th through 92nd percentiles), and high average (49th through 64th percentiles). The ability percentiles used here correspond to the percentiles forming AFQT categories.

Table 1 summarizes the 12 samples of aberrant response vectors. Each of these 24,000 (=12 × 2,000) response vectors was independently generated.

Analysis. All the item and test statistics required to compute the practical appropriateness indices were computed using the test norming sample. These quantities were computed as the first step in the analysis and then used in all subsequent analyses. LOGIST (Wood, Wingersky, & Lord, 1976) was used to estimate item parameters and a FORTRAN program was written to compute the other quantities required.

Then the 11 practical appropriateness indices were computed for the 4,000 response vectors in the normal (responding appropriately) sample. The item and test statistics estimated from the test norming sample were used in these calculations. This procedure simulates the process by which practical appropriateness indices would be computed in many applications. Four optimal indices were also computed for the normal sample: 15% spuriously high, 30% spuriously high, 15% spuriously low, and 30% spuriously low. The ability density f used in Equations 1 and 2 was the normal (0,1) density truncated to the interval [-2.05, 2.05]. The histograms used to generate the data were also used to compute the optimal indices; that is, polychotomous option characteristic curves were not estimated. (In order for an optimal index to be truly optimal for the corresponding form of aberrance, it is necessary to use the true option characteristic curves.)

Table 1. Ability Distributions Used to Generate Aberrant Samples

Percent of aberrant responses	Type of aberrance	
	Spuriously high	Spuriously low
15%	$N_T[-2.05,-1.34]$	$N_T(1.41,2.05]$
15%	$N_T(-1.34,-0.52]$	$N_T(0.35,1.41]$
15%	$N_T(-0.52,-0.05]$	$N_T(-0.05,0.35]$
30%	$N_T[-2.05,-1.34]$	$N_T(1.41,2.05]$
30%	$N_T(-1.34,-0.52]$	$N_T(0.35,1.41]$
30%	$N_T(-0.52,-0.05]$	$N_T(-0.05,0.35]$

Note. $N_T(a,b]$ is used to denote the standard normal distribution truncated to the interval (a,b]. Parentheses are used to indicate interval endpoints that were not included in the interval and brackets are used to indicate interval endpoints that were included in the interval.

The 11 practical appropriateness indices were computed for each of the 12 aberrant samples. In addition, the 15% spuriously high optimal index was computed for the three samples with this form of aberrance; the 30% spuriously high optimal index was computed for the three samples with this form of aberrance; etc.

Note that the ability density used in Equations 1 and 2 does not match the ability density of any aberrant sample. The proper interpretation of the optimal index is the following: It is the optimal index for the specified form of aberrance, say 15% spuriously high, in a population where the ability density is normal (0,1) truncated to [-2.05, 2.05] for both the normal and aberrant populations and a response vector is either normal or 15% spuriously high. The normal group does in fact have this ability distribution. By restricting the abilities of the aberrant group to a subinterval of [-2.05, 2.05], we determined the power in a particular subpopulation of the index that is optimal for the population as a whole.

Evaluation Criteria. The main criteria used for evaluating the appropriateness indices were the proportions of aberrant response patterns that were correctly identified as aberrant when various proportions of normal response patterns were misclassified as aberrant. These proportions were determined for all 12 aberrance conditions. This allowed us to determine what types of aberrant response patterns had acceptably high detection rates using optimal methods and using practical methods. The characteristics of response patterns that cannot be detected became evident as a consequence of examining the 12 aberrance conditions separately.

Results

Before presenting the results for the 12 aberrant samples, we shall illustrate some problems caused by poorly standardized appropriateness indices. Table 2 presents detection rates for the 15% spuriously high aberrant sample for two different samples of normal responses. In one case, the normal sample consists of the 200 response vectors with the highest θ values from the normal sample of $N = 4{,}000$ previously described; in the other case, the normal sample consists of the 200 response vectors with the lowest θ value. (Results for B are not given because this index was not programmed in its final form when this table was constructed.)

As shown in Table 2, the IOV index seems to be fantastic when the normal group consists of high ability normals: It correctly identified every single aberrant response vector, without a single misclassification of a normal. The S index appeared to be an excellent index, although not as powerful as IOV. In contrast, F1 seemed to be an abysmally poor index.

These results were almost completely contradicted for the low ability normals. At a 1% false alarm rate, the detection rate of the IOV index was 10% when the normal group consisted of low ability response patterns; it was 100% when the normals were high ability. The comparable rates for S were 78% and 8%, respectively. The results for F1 were in the opposite direction: The detection rate was 0% for normals of high ability but 34% for normals of low ability.

Table 2. Selected ROC Curve Points for the 15% Spuriously High Treatment, Aberrant Response Patterns Generated from 0-9% Ability Range

False alarm rate	Proportion detected by									
	z_ℓ	F1	F2	S	T2	T4	IOV	O/E	JK	NI
Normal Group = 200 High Ability Normals										
001	38	00	12	62	59	18	1.00	18	13	00
005	44	00	16	76	63	39	1.00	19	14	00
01	47	00	34	78	72	40	1.00	21	15	00
02	56	00	42	82	75	60	1.00	26	21	00
03	61	00	53	86	83	65	1.00	30	22	00
04	67	00	54	87	84	69	1.00	40	26	00
05	73	00	57	91	84	69	1.00	40	30	00
07	77	01	65	93	89	79	1.00	46	30	00
10	79	07	74	96	93	82	1.00	54	35	00
Normal Group = 200 Low Ability Normals										
001	26	25	25	00	14	21	00	00	00	00
005	31	33	27	06	38	26	05	00	00	00
01	44	34	36	08	44	30	10	01	03	00
02	48	46	42	11	49	32	13	03	05	00
03	50	48	46	11	50	37	16	04	09	00
04	52	49	54	20	53	45	18	06	12	00
05	61	54	57	24	59	48	23	09	17	00
07	67	63	61	30	64	53	29	11	19	00
10	72	69	67	35	76	62	33	18	23	00

Note. z_ℓ = standardized ℓ_0; F1 = fit statistic 1; F2 = fit statistic 2; T2 = second standardized extended caution index; T4 = fourth extended standardized caution index; IOV = item-option variance; O/E = observed information divided by expected information; JK = normalized jackknife estimate of variance; NI = number of Newton-Raphson iterations

The differences in detection rates for F1, S, and IOV resulted from their poor standardizations. In contrast, the well-standardized z_l had detection rates of 47% and 44% at a 1% misclassification rate. F2 also had similar detection rates: 34% and 36%. T2 is not standardized as well as T4; however, the detection rates for T2 were higher than the rates for T4. O/E and JK had moderately dissimilar detection rates across the two sets of normals. Finally, the detection rates for NI were identical across conditions; unfortunately, the detection rates were exceedingly poor.

The results for the 15% and 30% spuriously high samples for the low ability range (0th through 9th percentiles) are shown in Table 3. In this case, the normal group consisted of 4,000 response vectors that were generated from θ values sampled from the standard normal distribution truncated to [-2.05,2.05]. Note that the detection rates for z_l, F2, and T2 were fairly close to the rates for LR. It is clear from Table 5 that the 30% spuriously high treatment is very detectable: LR, z_l, and T2 all had detection rates of 90% or more when the error rate was 1%. Even the relative moderate 15% spuriously high treatment (which affected at most 13 items on the 85-item test) was fairly detectable: LR and z_l had detection rates of 50 and 46% at a 1% error rate. O/E and JK, which were shown to be well standardized in the previous section of this paper, had little power. At a 1% error rate, O/E and JK detected only 22% and 33% of the 30% spuriously high response vectors.

Table 4 presents the results for the 15% and 30% spuriously high treatment applied to the moderately low ability range (10th through 30th percentiles). It should be more difficult to detect aberrant response vectors in this ability range than in the low ability range because the expected number of responses changed due to the aberrance manipulation is smaller. Surprisingly, the detection rates for LR did not decrease sharply: At a 1% error rate, the detection rates were 50% versus 45% for 15% spuriously high, and 93% versus 89% for 30% spuriously high. The detection rates declined more rapidly for z_l (46% vs. 30% for 15% spuriously high; 90% vs. 75% for 30% spuriously high) and F2 (34% vs. 21%; 85% vs. 73%). The rates of decline of T2 and T4 were intermediate. T2 declined from 37% to 33% for 15% spuriously high and from 91% to 81% for the 30% treatment. T4 declined from 30% to 25% and from 87% to 79% .

The trends seen in Tables 3 and 4 continue in Table 5, which presents the results for the 15% and 30% spuriously high treatments applied to the low average ability range (31st to 48th percentiles). As shown in Table 5, the LR index provided detection rates that are roughly 50% higher than those of the best practical indices. For example, at a 1% error rate LR had a detection rate of 34% for the 15% treatment; z_l, F2, T2, T4 had detection rates of 18%, 15%, 23%, and 20%, respectively. The detection rates were 78% versus 51%, 53%, 51% and 57% for the 30% spuriously high condition at a 1% error rate.

Table 6 presents the results for the 15% and 30% spuriously low treatment applied to the high average ability sample (between the 49th and 64th percentiles). It is evident that the practical appropriateness indices are quite ineffective relative to the optimal index. At a 1% error rate, LR had a 47% detection rate for the 15% treatment; the highest rate for any of the practical indices was only 16%. The pattern of results for the 30% condition was similar. Here the LR detection rate was an impressive 79% when the error

Table 3. Selected ROC Curve Points for Aberrant Response Patterns Generated from the 0-9% Ability Range

False alarm rate	Proportion detected by											
	LR	z_p	F1	F2	S	T2	T4	IOV	O/E	JK	B	NI
15% Spuriously High												
001	30	26	00	12	10	13	13	13	00	00	00	00
005	43	40	00	27	31	25	21	29	01	02	00	00
01	50	46	00	34	45	37	30	42	04	06	00	00
02	59	54	08	44	59	50	41	53	11	12	02	00
03	64	60	22	51	67	56	49	63	16	17	05	00
04	67	64	32	55	72	62	54	70	20	21	09	00
05	70	69	40	60	78	66	59	75	24	23	13	00
07	73	74	52	69	83	73	65	82	31	30	21	00
10	77	80	63	76	89	81	73	89	42	39	35	00
30% Spuriously High												
001	85	78	00	63	22	75	69	20	01	07	00	00
005	91	87	01	81	51	86	80	37	09	17	00	00
01	93	90	11	85	65	91	87	50	22	33	00	00
02	95	93	44	90	79	95	92	60	43	49	00	00
03	95	95	69	93	85	96	94	70	53	56	00	00
04	96	96	80	94	88	97	95	76	57	65	00	00
05	97	96	86	95	92	97	96	80	62	65	00	00
07	97	97	92	97	94	98	97	86	72	72	01	00
10	98	98	95	98	96	98	98	91	80	78	04	00

Note. The maximum detection rate among the reasonably well-standardized indices is underlined at each false alarm rate.

Table 4. Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 10-30% Ability Range

False alarm rate	Proportion detected by											
	LR	z_t	F1	F2	S	T2	T4	IOV	O/E	JK	B	NI
15% Spuriously High												
001	23	14	00	06	01	13	11	00	00	00	00	00
005	37	23	00	16	05	22	17	02	01	02	00	00
01	45	30	00	21	10	33	25	05	03	05	00	00
02	55	38	05	31	19	44	36	09	08	11	00	00
03	60	45	15	38	25	49	43	13	12	15	00	00
04	63	49	22	43	30	53	47	17	15	19	00	00
05	66	53	28	47	38	57	51	21	18	21	00	00
07	70	59	41	56	46	64	58	30	26	27	01	00
10	75	65	52	63	58	71	66	40	35	35	03	00
30% Spuriously High												
001	76	56	00	45	04	61	60	01	08	17	00	00
005	85	71	04	67	15	72	72	04	22	29	00	01
01	89	75	11	73	27	81	79	08	35	43	00	01
02	92	82	34	81	40	87	86	13	52	58	00	01
03	93	86	57	86	49	90	90	18	61	64	00	01
04	94	88	68	88	56	92	92	22	65	69	00	01
05	95	90	75	90	64	93	93	26	70	72	00	01
07	96	92	83	93	71	94	95	35	77	77	00	01
10	97	94	88	95	80	96	96	45	84	84	00	01

Table 5. Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 31-48% Ability Range

False alarm rate	Proportion detected by											
	LR	z_1	F1	F2	S	T2	T4	IOV	O/E	JK	B	NI
15% Spuriously High												
001	13	07	00	04	00	09	08	00	01	02	00	00
005	26	13	00	12	00	15	14	00	05	05	00	00
01	34	18	01	15	01	23	20	00	08	10	00	00
02	46	24	06	23	03	32	29	00	16	17	00	00
03	51	31	13	29	05	37	35	00	21	22	00	00
04	55	34	19	33	07	42	39	01	25	26	00	00
05	58	38	25	37	12	45	44	01	29	28	00	00
07	64	44	33	45	17	51	50	02	36	34	00	00
10	70	52	42	53	26	58	57	05	43	41	00	00
30% Spuriously High												
001	59	31	01	31	00	30	38	00	11	20	00	00
005	72	45	08	47	03	41	49	00	26	31	00	03
01	78	51	15	53	07	51	57	00	38	44	00	03
02	84	59	29	63	14	59	67	00	53	58	00	03
03	87	65	44	69	19	64	72	01	60	63	00	03
04	89	68	50	72	23	68	76	01	64	67	00	03
05	91	72	56	75	30	72	79	02	68	70	00	03
07	93	77	64	81	39	76	82	04	74	75	00	03
10	95	82	71	85	49	81	87	07	79	80	00	03

Table 6. Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 49-64% Range

False alarm rate	Proportion detected by											
	LR	z,	F1	F2	S	T2	T4	IOV	O/E	JK	B	NI
15% Spuriously Low												
001	29	06	00	03	00	04	04	00	00	01	00	00
005	43	12	01	08	00	08	07	00	01	02	00	00
01	47	16	03	11	00	14	11	00	03	06	00	00
02	56	22	09	17	02	20	17	01	09	12	00	00
03	61	27	17	21	03	24	21	02	12	17	00	00
04	63	30	24	25	05	28	26	04	15	20	00	00
05	67	35	29	29	08	32	29	06	18	23	00	00
07	71	40	37	37	13	38	35	10	23	29	00	00
10	76	49	46	44	20	46	42	17	32	37	00	00
30% Spuriously Low												
001	56	19	00	09	00	09	12	01	00	01	00	00
005	75	29	00	20	02	14	20	07	02	05	00	00
01	79	35	01	26	06	23	28	14	07	14	00	00
02	86	44	08	36	15	32	38	22	20	27	00	00
03	89	51	18	42	22	37	45	30	26	33	00	00
04	91	55	26	47	27	42	50	37	31	40	00	00
05	93	59	34	52	35	47	55	42	36	43	01	00
07	95	64	44	60	45	53	60	54	46	50	02	00
10	97	70	56	66	56	60	67	66	57	59	05	00

Table 7. Selected ROC Curve Points for the Aberrant Response Patterns Generated from the 65-92% Ability Range

False alarm rate	Proportion detected by											
	LR	z_1	F1	F2	S	T2	T4	IOV	O/E	JK	B	NI
15% Spuriously Low												
001	55	26	05	17	00	17	12	00	03	09	00	00
005	66	38	19	32	01	26	20	00	12	16	00	01
01	68	44	30	37	03	36	26	01	21	26	00	01
02	73	52	47	46	06	45	36	02	32	37	00	01
03	75	58	59	53	09	50	42	03	38	43	00	01
04	77	62	65	56	13	55	47	05	42	47	00	01
05	78	65	70	60	18	58	51	06	46	50	00	01
07	81	70	76	67	26	63	56	10	52	55	00	01
10	83	76	81	72	36	69	63	16	58	62	00	01
30% Spuriously Low												
001	80	54	00	40	01	44	45	04	04	12	00	00
005	89	66	08	58	09	54	55	13	15	27	00	00
01	91	71	18	62	19	64	63	24	31	44	00	00
02	94	78	42	72	32	74	72	32	48	59	00	00
03	95	83	59	77	40	77	77	41	55	64	00	00
04	96	85	69	80	47	80	80	47	61	71	00	00
05	97	87	75	83	55	83	82	53	67	74	00	00
07	98	89	82	87	63	86	86	63	75	80	00	00
10	98	92	88	91	74	90	89	72	81	85	00	00

Table 8. Selected ROC Curve Points for the Aberrant Response Patterns Generated from 93-100% Ability Range

False alarm rate	Proportion detected by											
	LR	z_1	F1	F2	S	T2	T4	IOV	O/E	JK	B	NI
15% Spuriously Low												
001	73	55	26	39	01	31	23	00	06	12	00	01
005	80	68	59	57	10	42	33	00	15	20	00	08
01	81	72	71	62	17	54	41	01	21	30	00	08
02	84	78	82	72	27	63	52	02	33	43	00	08
03	86	82	88	77	36	67	57	03	38	49	00	08
04	86	84	90	80	43	71	63	05	43	54	00	08
05	88	87	91	82	50	74	66	06	47	56	00	08
07	89	90	93	86	60	79	71	11	56	64	00	08
10	91	92	94	89	69	84	77	16	64	72	00	08
30% Spuriously Low												
001	93	88	06	78	10	83	79	09	27	47	00	00
005	96	93	38	88	32	90	86	21	53	65	00	03
01	97	95	59	91	47	94	90	31	68	78	00	03
02	98	97	81	94	63	96	94	41	82	88	00	03
03	98	98	92	96	72	97	95	51	87	90	00	03
04	98	98	95	97	76	98	96	59	89	93	00	03
05	99	98	96	98	82	98	97	63	91	94	00	03
07	99	98	98	98	88	98	98	72	94	96	00	03
10	99	99	98	99	93	99	98	80	96	97	00	03

rate was 1%; the next best index (z_3) detected only 35% of the aberrant sample.

In Table 7, which presents the results for the 15% and 30% spuriously low samples with θs in the 65th through 92nd percentiles, the practical appropriateness indices have detection rates that are closer to the rates of the optimal index. This trend is continued in Table 8, which presents the results for the spuriously low treatments applied to the highest ability category (percentiles 93 and above). At a 1% error rate, for example, LR detected 81% of the 15% spuriously low response patterns; z_3, F2, and T2 had detection rates of 72%, 62%, and 54%. For the 30% treatment, the rate for LR was 97%; z_3, F2, and T2 had rates of 95%, 91%, and 94%.

Drasgow and Guertler (1987) recently presented a utility theory approach to the use of Appropriateness Measurement in practical settings. Their approach requires the densities of an index in normal and aberrant samples. Consequently, normal distributions were fitted to the distributions of z_3, F2, and T4 by equating the first two moments of the normal distribution to the empirical moments. These analyses were based on the first 1,000 response vectors from the normal sample and each of the 12 aberrant samples. The fitted means and standard deviations are presented in Table 9. As a crude measure of fit, Kolmogorov-Smirnov test statistics were computed to compare the empirical distributions to normal distributions with the observed moments. No significant ($\alpha = .05$) departures of empirical distributions from the corresponding fitted normal distribution were found. As the Kolmogorov-Smirnov test can be conservative when fitted moments are substituted into the theoretical distribution (Massey, 1951), these results should be viewed with some caution.

Discussion

There has been a growing interest in Appropriateness Measurement, both by researchers and by testing practitioners. To date, however, there has been little critical study of the various indices available. The results of the research summarized here clearly indicate that there are important differences in the properties of appropriateness indices. Figures 1 through 3 show that some indices are poorly standardized (e.g., IOV), and a "standardized" index may not be well standardized (e.g., F1). Table 2 illustrates the problems that are caused by poorly standardized indices.

A well-standardized index is not, however, necessarily a good appropriateness index. The O/E and JK indices were shown to be reasonably well standardized in Figures 1 through 3, but Tables 3 through 8 clearly show them to be ineffective in detecting aberrant response patterns.

Perhaps the most important finding of the simulation reported in this chapter is that z_3, F2, and T2 provide nearly optimal rates of detection of some forms of aberrance but inadequate rates of detection of other forms of aberrance. In particular, these three indices have near-optimal rates of detection when the spuriously high treatment is applied to very low ability response vectors and when the spuriously low treatment is applied to very high ability response vectors. Unfortunately, these indices have rates of detection far below optimal when the spuriously high and low treatments are applied to response vectors with nearly average ability values.

Table 9. Means and Standard Deviations of Empirical Distributions of z_3, F2, and T4

Aberrance manipulation	Ability range	Severity of aberrance: 15% z_3	15% F2	15% T4	30% z_3	30% F2	30% T4
Spur. High	0-9%	-2.32	1.28	1.56	-4.00	1.49	3.22
		(1.13)	(0.14)	(0.94)	(1.22)	(0.15)	(1.07)
Spur. High	10-30%	-1.85	1.23	1.39	-3.32	1.43	3.04
		(1.11)	(0.14)	(0.98)	(1.19)	(0.15)	(1.10)
Spur. High	31-48%	-1.38	1.19	1.22	-2.47	1.36	2.38
		(1.03)	(0.14)	(1.02)	(1.21)	(0.17)	(1.19)
Spur. Low	49-64%	-1.02	1.13	0.65	-1.58	1.19	1.20
		(1.03)	(0.13)	(0.99)	(1.14)	(0.14)	(0.98)
Spur. Low	65-92%	-1.85	1.23	1.17	-2.74	1.34	2.12
		(1.16)	(0.16)	(1.11)	(1.19)	(0.15)	(1.08)
Spur. Low	93-100%	-3.01	1.37	1.78	-4.28	1.54	3.50
		(1.30)	(0.17)	(1.14)	(1.32)	(0.17)	(1.24)
Normals[a]	0-100%	0.09	0.99	-0.14			
		(0.97)	(0.12)	(0.86)			

Note. Means and standard deviations are based on samples of N = 1000. Standard deviations are in parentheses.

[a]To conserve space, results for the normal sample are listed under the columns for the 15% severity of aberrance.

These results indicate that we need to devise new indices that are more powerful than z_l, F2, and T2 for examinees whose abilities are near average. We expect that it may be necessary to construct two indices: one for spuriously low response patterns and one for spuriously high response patterns. This psychometric necessity would be quite useful for practitioners because it would allow them to diagnose the cause of aberrance in addition to detecting aberrant response patterns.

III. POLYCHOTOMOUS ANALYSIS OF THE ARITHMETIC REASONING TEST: AN APPLICATION OF MULTILINEAR FORMULA SCORE THEORY

Introduction

Multilinear formula score theory or multilinear formula scoring (MFS; Levine, 1983, 1985a, 1985b) is a nonparametric IRT for which consistent and asymptotically efficient estimators of ability densities, item characteristic curves (ICCs), and option characteristic curves (OCCs) have been derived and programmed. MFS provides a powerful new approach to substantive questions of long standing. These questions include determining the shapes of ability distributions and the magnitudes of differences among ability distributions of various groups, determining the shapes of item characteristic curves for unidimensional and multidimensional tests, identifying biased and other faulty items, and assessing the extent to which two tests measure the same ability.

In the research reported this chapter, we used three-parameter logistic ICCs to model the way in which examinees respond to correct options of AR multiple-choice items and, simultaneously, we used MFS to model responses to the incorrect options. Thus, we replaced the crude "histogram model" of Chapter II with a theory-based approach. Consequently, low rates of detection of inappropriate response patterns cannot be attributed to an unsophisticated analysis of the data.

Prior to determining rates of detection of spuriously high and low response patterns, we examined MFS's ability to estimate option response curves. The results of this analysis were assessed graphically and by determining the increase in information about ability due to polychotomous scoring of item responses. The term "information" is used in its statistical sense to mean the expected squared derivative of the logarithm of the likelihood function. Since the asymptotic standard error of the maximum likelihood estimate of an ability θ equals the square root of the reciprocal of the information function at θ, an increase in information due to polychotomous scoring is readily translated into percent test length reduction made possible by polychotomous scoring.

We also compared the dichotomous and polychotomous item response models' potentials for supporting Appropriateness Measurement. Of course, the model-based detectability of a particular type of aberrance depends upon the item response model used to analyze the data; more specific (polychotomous) models are expected to be rejected more frequently when fitted to aberrant response patterns and thus provide superior appropriateness measurement. By combining the optimal appropriateness index results of Levine and Drasgow (1984, 1987)

with MFS's ability to accurately recover the option characteristic curves needed for polychotomous modeling, we determined whether polychotomous modeling was negligibly or markedly superior to dichotomous modeling in detecting test anomalies.

This chapter also contributes to formula score theory in that it provides a verification of MFS theoretical results with simulation data.

Review of Multilinear Formula Score Theory

This section contains a review of MFS theory as it is used in this paper. The theory is more general than outlined here, but for the sake of clarity, we will describe only the special case required for the present research.

Let $\underline{u}_i$ denote the response to the $\underline{i}$th item of an $\underline{n}$ item test scored $\underline{u}_i = 1$ if correct and $\underline{u}_i = 0$ if incorrect. The $\underline{u}_i$ generate the elementary formula scores, which can be enumerated as

$$
\begin{aligned}
&1 \\
&\underline{u}_1, \underline{u}_2, \cdots, \underline{u}_n \\
&\underline{u}_1\underline{u}_2, \underline{u}_1\underline{u}_3, \cdots, \underline{u}_{n-1}\underline{u}_n \\
&\cdots \\
&\underline{u}_1\underline{u}_2 \cdots \underline{u}_n \ .
\end{aligned}
$$

Traditional formula scoring (Lord & Novick, 1968, Chapter 14) generally uses only linear scores. When there is neither omitting nor polychotomous scoring, linear formula scores are formulas with a constant term plus a linear combination of the binary item scores, $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_n$. (When there is omitting and polychotomous scoring, a linear score is a constant plus a linear combination of binary variables indicating omitting and option choice.)

Multilinear formula score theory generalizes traditional formula score theory by using quadratic scores (linear scores added to linear combinations of $\underline{u}_1\underline{u}_2, \underline{u}_1\underline{u}_3, \ldots, \underline{u}_{n-1}\underline{u}_n$), cubic scores (quadratic scores plus linear combinations of products of item scores for three different items), and higher order scores. Most of the results in this chapter were obtained with fifth order scores. The new theory is called "multilinear" because frequent use is made of the fact that when all the scores except one are held constant, a "linear" score is obtained.

In this chapter, as in Chapter II, we assume that the regression of $\underline{u}_i$ on the latent trait θ is a three-parameter logistic ogive. By local independence, the regressions of the elementary formula scores on the latent trait can then be written as

$$
\begin{aligned}
&1 \\
&\underline{P}_1(\underline{t}), P_2(\underline{t}), \ldots, \underline{P}_n(\underline{t}) \\
&\underline{P}_1(\underline{t})P_2(\underline{t}), \underline{P}_1(\underline{t})\underline{P}_3(\underline{t}), \ldots, \underline{P}_{n-1}(\underline{t})\underline{P}_n(\underline{t}) \\
&\cdots
\end{aligned}
$$

$$\underline{P}_1(\underline{t})\underline{P}_2(\underline{t}) \ \ldots \ \underline{P}_n(\underline{t}) \ ,$$

where $\underline{t}$ is used to denote a specific value of θ.

There are 2^n regression functions listed above. More can be generated by taking linear combinations of the elementary formula scores and then computing their regressions on the latent trait. For example, the number-right score

$$\underline{X} = \underline{u}_1 + \underline{u}_2 + \ldots + \underline{u}_n$$

has the regression

$$E(\underline{X} \mid \underline{t}) = \sum_{i=1}^{n} \underline{P}_i(\underline{t}) \ .$$

The collection of regression functions of all linear combinations of elementary formula scores is called the canonical space (CS) of a test.

A major step in an MFS analysis of a test consists of finding a smaller number of functions to represent the large number (in fact, an infinite number) of functions in the canonical space. The smaller collection of functions is called an orthonormal basis for the canonical space.

Selecting an orthonormal basis for the canonical space is analogous to finding the principal components of a set of variables. In a principal components analysis, the basic idea is to create a new set of variables, the principal components, so that each of the original variables can be written as a linear combination of the principal components plus a small residual. A principal components analysis is valuable when there is a large number of original variables and the first few principal components explain almost all of their variance. In the same way, functions in the canonical space are written as linear combinations of the orthonormal basis functions. For example, the ICC for the $\underline{i}$th item can be written

$$\underline{P}_i(\underline{t}) = \sum_{k=1}^{K} a_k \underline{h}_k(\underline{t}) \ ,$$

where $\underline{K}$ functions, denoted $\underline{h}_1(\underline{t}), \ldots, \underline{h}_K(\underline{t})$, are used in the orthonormal basis and the a_k are the weights used in the linear combination. If $\underline{K}$ is sufficiently large, this representation is exact. If only the first $\underline{J}$ functions are used, instead of all $\underline{K}$ functions (where $\underline{J}$ is less than $\underline{K}$), then there is some error. However, the residual

$$\underline{P}_i(\underline{t}) - \sum_{k=1}^{J} a_k \underline{h}_k(\underline{t}) = \sum_{J+1}^{K} a_k \underline{h}_k(\underline{t})$$

will be small if the α_k are small for values of k larger than J. In fact, the area under the squared residual is exactly $\alpha^2_{J+1} + \alpha^2_{J+2} + \ldots + \alpha^2_K$.

In each MFS analysis, a parsimonious representation of one or another collection of functions in the CS is important. MFS provides techniques that yield basis functions that give small values of α_k for large values of k, at least for the collection of functions being analyzed. Most MFS analyses require six to eight basis functions for an adequate representation of the functions being studied.

To recapitulate, the analysis begins by estimating ICCs from the dichotomously scored item responses. Widely available programs such as LOGIST (Wood, Wingersky, & Lord, 1976) and BILOG (Mislevy & Bock, 1983) can be used to this end. The estimated ICCs and the assumption of local independence are subsequently used to define the canonical space. Then a small number of orthonormal basis functions are selected so that the functions in the canonical space are well approximated by linear combinations of the orthonormal basis functions.

The next step of the MFS analysis is to determine weights for the orthonormal basis functions so that option characteristic curves (OCCs) can be written as linear combinations of the h_js. Since OCCs were not included in the set of functions used to define the canonical space, we must address both the mathematical question of how best to approximate the OCCs by basis functions and the substantive question of whether or not the basis functions can adequately approximate OCCs. The OCC analysis proceeds item-by-item, with the weights for all the options (including omit as an option) to each item simultaneously estimated by the method of marginal maximum likelihood. The log likelihood that is maximized with respect to the weights is

$$L = \sum_{j=1}^{N} \log P(u_j, v_{ij}) , \qquad (17)$$

where u_j is a vector containing the dichotomously scored item responses of the jth examinee and v_{ij} indicates the particular option on item i selected by examinee j. For a four-option multiple-choice item, $v_{ij} = 1$ if option A is selected, ..., $v_{ij} = 4$ if option D is selected, and $v_{ij} = 5$ if no response is made. Suppose all the items are recoded such that option A is always the correct response. Then Equation 17 can be rewritten as

$$L = \sum_{\substack{j=1 \\ v^*_{ij}=1}}^{N} \log P(u_j) +$$

$$\sum_{\substack{j=1 \\ v_{ij}^{*} \neq 1}}^{N} \log \int \underline{P}(u_j \mid \underline{t})\, \underline{P}(\underline{v}_{ij} \mid \underline{t}, \underline{u}_{ij}=0)\underline{f}(\underline{t})dt \qquad (18)$$

where

$$\underline{P}(u_j \mid \underline{t}) = \prod_{i=1}^{n} \underline{P}_i(\underline{t})^{u_{ij}}[1 - \underline{P}_i(\underline{t})]^{1-u_{ij}}, \qquad (19)$$

$$\underline{P}(\underline{v}_{ij} \mid \underline{t}, \underline{u}_{ij} = 0) = \sum_{k=1}^{J} \alpha_k \underline{h}_k(\underline{t}), \qquad (20)$$

and $\underline{f}(\underline{t})$ is the ability density. Notice that Equation 19 is the likelihood function for the three-parameter logistic model [i.e., Lord's (1980) Equation 4-20 and Hulin et al.'s (1983) Equation 2.6.2]. It is the α_ks in Equation 20 that are to be estimated. Actually, each option has its own set of $\underline{J}$ α_ks, but to avoid notational complexity, we have not added another subscript to the α_ks.

It is important to observe that local independence is not used to derive Equation 18 from Equation 17; only the definition of conditional probability is used. Thus, even when skipping items or not reaching items (response "5") fails to obey the assumption of local independence, an accurate estimate of the conditional probability of non-response for examinees at each ability level is obtained.

Quadratic programming methods are used to obtain maximum likelihood estimates of orthonormal basis function weights for conditional option characteristic curves (COCCs) in Equation 20. A COCC equals its associated OCC divided by $[1-\underline{P}_i(\theta)]$; hence, the COCCs for an item sum to 1 for all θ values. The OCCs for an item, in contrast, sum to $[1-\underline{P}_i(\theta)]$, which becomes very small as $\underline{P}_i(\theta)$ approaches 1. The weights α_k for the COCCs are easier to estimate than the weights for OCCs since the OCCs for easy items and for rarely chosen options are close to 0, which causes the α_k to become indeterminant; COCCs are not usually close to 0. Because the OCC at $\theta = \underline{t}$ is equal to the COCC multiplied by $1 - \underline{P}_i(\underline{t})$, the OCCs are available after the COCCs have been obtained. The COCCs are intrinsically interesting as well as mathematically tractable since their shapes can be used to study the properties of effective distractors.

The quadratic programming methods used by Levine and Williams (1985) are convenient because they allow plausible constraints to be placed on the COCCs. One constraint is positivity: COCCs are not allowed to become negative. In the present analyses all COCCs were required to equal or exceed .001. A second constraint placed on COCCs is smoothness: The COCCs were not allowed to oscillate widely. The smoothness constraint was implemented by restricting

the third derivative of the COCCs to be less than .005. This condition can be thought of as requiring each small piece of the graph of the COCC to have a very accurate quadratic approximation. (A restriction on the second derivative would force the COCC to be locally linear, and a first derivative constraint would force the COCC to be locally constant.)

Estimation and Information

Data set. The data set used in our analyses was a spaced sample of 2,978 examinees taken from the National Opinion Research Center (NORC; Bock & Mislevy, 1981) sample of American youths. These examinees answered the 30-item ASVAB Arithmetic Reasoning (AR) subtest. Each item on this test has four options.

ICC estimation. The first step in the MFS analysis was to estimate ICCs from the dichotomously scored item responses. To this end, the item responses of the examinees described above were scored dichotomously. All nonanswered items were scored as incorrect (since we treated omits as a separate--and incorrect--response option). Then version 2B of LOGIST (Wood et al.,1976) was used to estimate item and ability parameters. Estimates of item discrimination parameters ranged from about 0.5 to 2.0, and estimates of item difficulties varied from about -3.0 to 1.4 (mean = .14, SD = .99).

Density estimation. The ability density f shown in Equation 18 was estimated by the nonparametric method developed by Levine and Williams (1985). The density was represented as a linear combination of basis functions, and the weights were estimated by maximum likelihood. The weight vectors were restricted to a convex set determined by hypotheses about the shape of the unknown density. After experimenting with various shape hypotheses, the following conditions were selected. The density was constrained to be nonnegative; to have a nonnegative second derivative between -4.8 and -3.1; to have a nonpositive second derivative for abilities between -.3 and 1.0; to be monotonically increasing for abilities between -3.1 and -.3; and to be monotonically decreasing for abilities between 1.0 and 3.5. These conditions imply that the density will be unimodal between -3.1 and 3.5, that the mode will occur between -.3 and 1.0, and that the density will either decrease to a lower asymptote as ability decreases to -5 or will have a second mode in the left tail if such is indicated by the data. It was decided to allow a second maximum at very low abilities because the data seemed substantially better fit when bimodality was permitted. A substantive interpretation of bimodality is noted below.

After some preliminary analyses, we decided to remove examinees who answered less than half of the items. There were 87 such examinees, leaving 2,891 examinees for the density and COCC estimation.

Figure 4 shows the obtained density. It can be seen that the density is roughly bell-shaped, with a mode near 0. The left tail turns up at low abilities, suggesting a relatively large number of examinees with very low abilities. One substantive interpretation of this fat left tail is that even among examinees who answered more than half of the items there may have been some who were poorly motivated and did not make a serious attempt to pass the examination. In fact, examinees were paid to take the examination and consequently some of them may not have been adequately motivated. The test

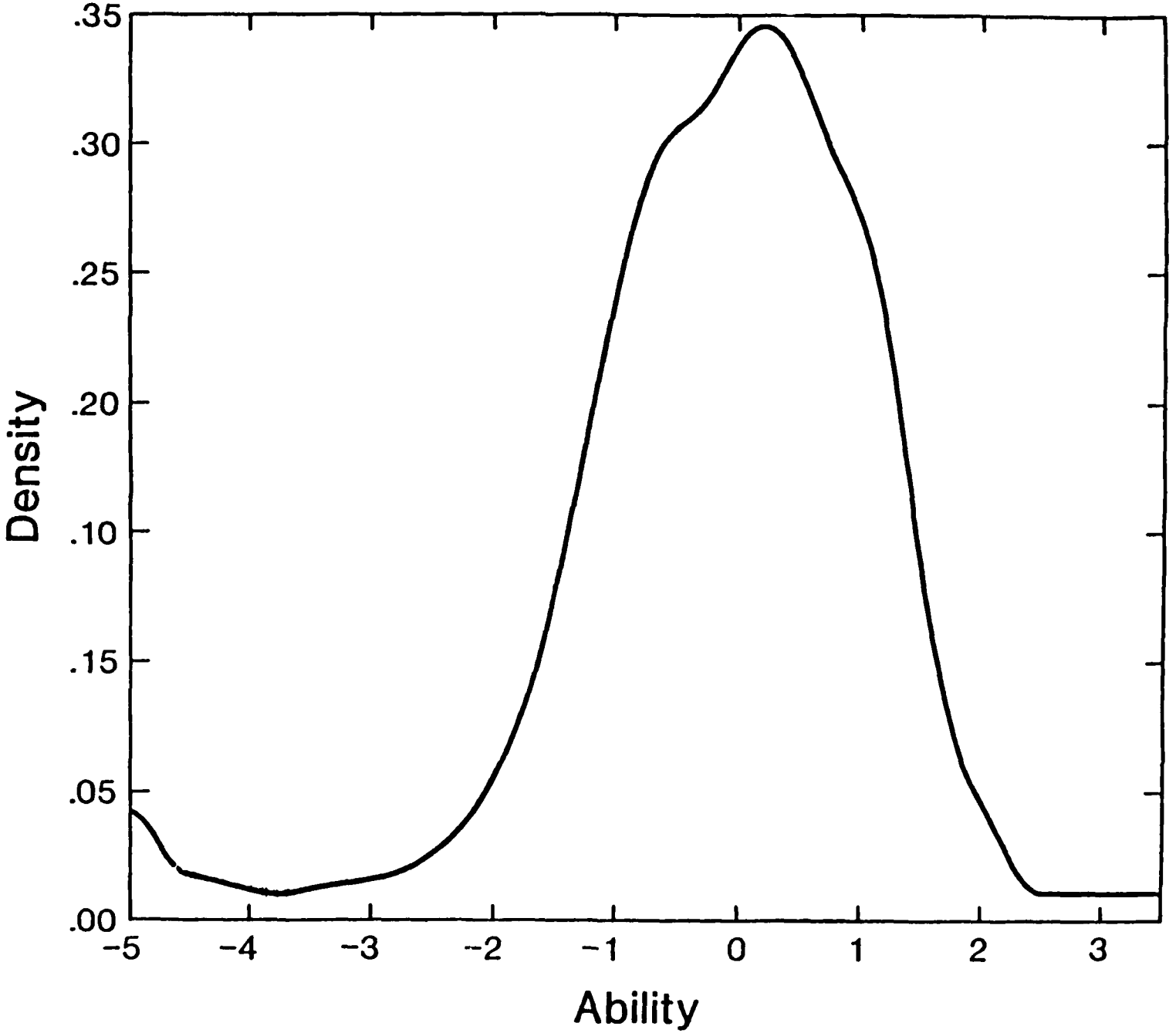


Figure 4. Ability density for National Opinion Research Center sample on the Arithmetic Reasoning subtest.

information function at $\theta = -5$ is very low; consequently, bimodality cannot be established unequivocally without much larger samples.

COCC estimation. Four COCCs were estimated for each item: the three incorrect response curves and an omit curve. Omits included both skipped responses and not-reached responses. The number of orthonormal basis functions used in the analysis was 10. Thus, 30 weights (10 weights for each of three COCCs) were estimated for each item. The weights for the fourth COCC were a known linear combination of the weights for the other three (Levine, 1985b).

Appendix A contains plots of the COCCs estimated for all 30 AR items. The solid curves indicate the estimated COCCs. Each page in Appendix A contains the four COCCs for two items. For example, the first page of Appendix A has the COCCs for item 1 plotted in the four panels to the left; the four panels to the right contain COCCs for Item 2 of the AR subtest. For each item, the top left panel contains the COCC for the first incorrect option; the top right panel, the COCC for the second incorrect option; the bottom left panel, the COCC for the third incorrect option; and the bottom right panel, the omit COCC.

The goodness-of-fit of the estimated COCCs can be evaluated by examining the vertical lines displayed in each panel. These lines were obtained by computing three-parameter logistic ability estimates for all 11,914 examinees in the NORC data set, forming 25 ability strata on the basis of estimated abilities by using the 4th, 8th, ..., 96th percentiles of the standard normal distribution as cutting scores, and then computing, from among the subset of examinees who answered the item incorrectly, the proportion of examinees selecting each option. The centers of the vertical lines correspond to the observed proportions and they are plotted above the category medians (the 2nd, 6th, ..., 98th percentiles of the standard normal distribution). The vertical lines represent approximate 95% confidence intervals for the observed proportions (± two standard errors, where the observed proportion is used to compute the standard error). Observed proportions of 0 and 1 are plotted as plus signs and are offset slightly from their true locations so that they will be visible.

The AR items seem to be more-or-less ordered by difficulty. Consequently, the 95% confidence intervals for the first few items in Appendix A are very wide because these items are easy and so few examinees chose incorrect options. Confidence intervals for later items are much narrower and provide a severe test for COCC estimates. Item 27, for example, shows that the COCC estimates provide a very good description of option choice. Notice that the COCC for the omit category lies below most observed proportions. This occurs because examinees with high omitting rates were excluded from the sample used to estimate COCCs, but were included in the total sample used to compute the proportions displayed in Appendix A.

COCC estimation verification. The figures presented in Appendix A show that MFS estimates of COCCs closely follow the actual patterns of item responses. It is difficult, however, to understand the accuracy of COCC estimates from these figures because the true COCCs are not known. To gain further insights into the properties of MFS estimates of COCCs, a simulation data set of 3000 response patterns was generated. Simulated abilities were

sampled from the standard normal distribution, probabilities of correct and incorrect responses were determined from the ICCs obtained by the LOGIST run described previously, and probabilities of option selections (for responses simulated to be incorrect) were computed using the MFS-estimated COCCs.

COCCs were re-estimated from the simulation data set. The true ability density (the standard normal) was used in Equation 18, and the true ICC values were used to compute probabilities of correct and incorrect responses. The true ability density and ICC values were used because we wanted to determine the errors of COCC estimates in a way that was not confounded with inaccuracies in density estimates and ICC estimates.

The results of the simulation study are shown in Appendix B, which presents the re-estimated COCCs for all 30 items. Heavy lines indicate the re-estimated COCCs and thin lines indicate the true COCCs. Observed proportions and their approximate 95% confidence intervals are shown for the simulation sample of $\underline{N}$ = 3,000. The observed proportions were not plotted if five or fewer incorrect responses were made in an ability stratum.

Item 2 shows estimated COCCs that are very close to the true COCCs for all ability levels. This is remarkable because there were almost no incorrect responses made by simulated examinees with above-average ability. Item 3 shows that we cannot always expect to have well-estimated COCCs when there are no data available: Large diffences between true and estimated COCCs occur at high ability levels. The COCCs were, however, accurately estimated in ability ranges for which there were more than a handful of incorrect responses.

From an inspection of the plots in Appendix B, it seems evident that COCC values were accurately estimated when there were six or more incorrect responses in adjacent ability strata. Sometimes COCC values were well-estimated when fewer incorrect responses were available, but this seemed to be a matter of chance. Notice, also, that COCCs for the omit option were not underestimated in this analysis as they were in the analysis of the real AR data. In this analysis, all response vectors were used; there was no restriction on omitting as in the previous analysis.

Information functions. Information functions for the dichotomous and polychotomous modelings of the AR test are shown in Figure 5. An expression for the information function of the three-parameter logistic model is

$$\text{Information at } \underline{t} = \sum_i \frac{[\underline{P}_i'(\underline{t})]^2}{\underline{P}_i(\underline{t})} + \sum_i \frac{[\underline{Q}_i'(\underline{t})]^2}{\underline{Q}_i(\underline{t})} \qquad (21)$$

where $\underline{Q}_i = 1 - \underline{P}_i$ and $\underline{P}_i'$ and $\underline{Q}_i'$ are the first derivatives of $\underline{P}_i$ and $\underline{Q}_i$. The information function of the polychotomous model is

$$\text{Information at } \underline{t} = \sum_i \frac{[\underline{P}_1'(\underline{t})]^2}{\underline{P}_1(\underline{t})} + \sum_i \sum_{j=2}^{J} \frac{[\underline{P}_{ij}'(\underline{t})]^2}{\underline{P}_{ij}(\underline{t})} \qquad (22)$$

where P_{ij} is the OCC for option j on item i and P'_{ij} is its first derivative. The correct option makes the same contribution to information for both the dichotomous and polychotomous scorings; namely, the first term on the right sides of Equations 21 and 22. Thus, any differences in information are entirely due to the treatment of incorrect responses. Although it is not obvious from Equations 21 and 22, it can be shown that the information function for the polychotomous model equals or exceeds the three-parameter logistic model's information function. Thus, any increase in information is entirely due to polychotomous scoring.

Figure 5 shows that there are moderate gains in information due to polychotomous scoring of the AR items for low to moderately high abilities. These gains are equivalent to adding about 5 or 6 items to the subtest. Little or no information is gained for high ability examinees. This latter finding is not surprising because high ability examinees are expected to answer nearly all the items correctly.

It should be noted that the AR items were not written with polychotomous scoring in mind, and so the gains in information shown in Figure 5 are more-or-less fortuitous. Larger gains might be realized if item writers knew the attributes of incorrect options that typically lead to substantial increases in information.

Appropriateness Measurement for the AR Subtest

Purpose

This section compares the effectivenesses of dichotomous and polychotomous models for detecting aberrant responses patterns. By comparing detection rates of optimal indices, it is possible to compare the maximum detection rates possible for a given form of aberrance. As in the previous section, the dichotomous model is a submodel of the polychotomous model; hence, any increase in detection rates is due to modeling incorrect responses.

Several practical indices were also evaluated. Most of these indices are computed from the dichotomously scored item responses. One index, however, is the natural extension of a dichotomous model index to the polychotomous case. Detection rates for the practical indices will indicate (a) which are relatively more powerful and less powerful, and (b) the extent to which the maximum detection rates are attained.

Overview

The ICCs and OCCs estimated for the AR subtest from the sample of $N = 2,891$ were used as the "true" item parameters in a simulation study. Initially, a sample of $N = 3,000$ simulated response patterns was created and used as a test norming sample. This data set was used to determine the item and test statistics required to compute all but two (z_p and DFK) of the practical appropriateness indices listed in the next section. Then a normal sample (appropriate responding) of $N = 4,000$ response vectors was created. In addition, 16 aberrant samples of $N = 2,000$ were generated to simulate several forms of aberrance. Optimal indices and all the practical indices were then computed for the normal sample and aberrant samples. Rates of detection of

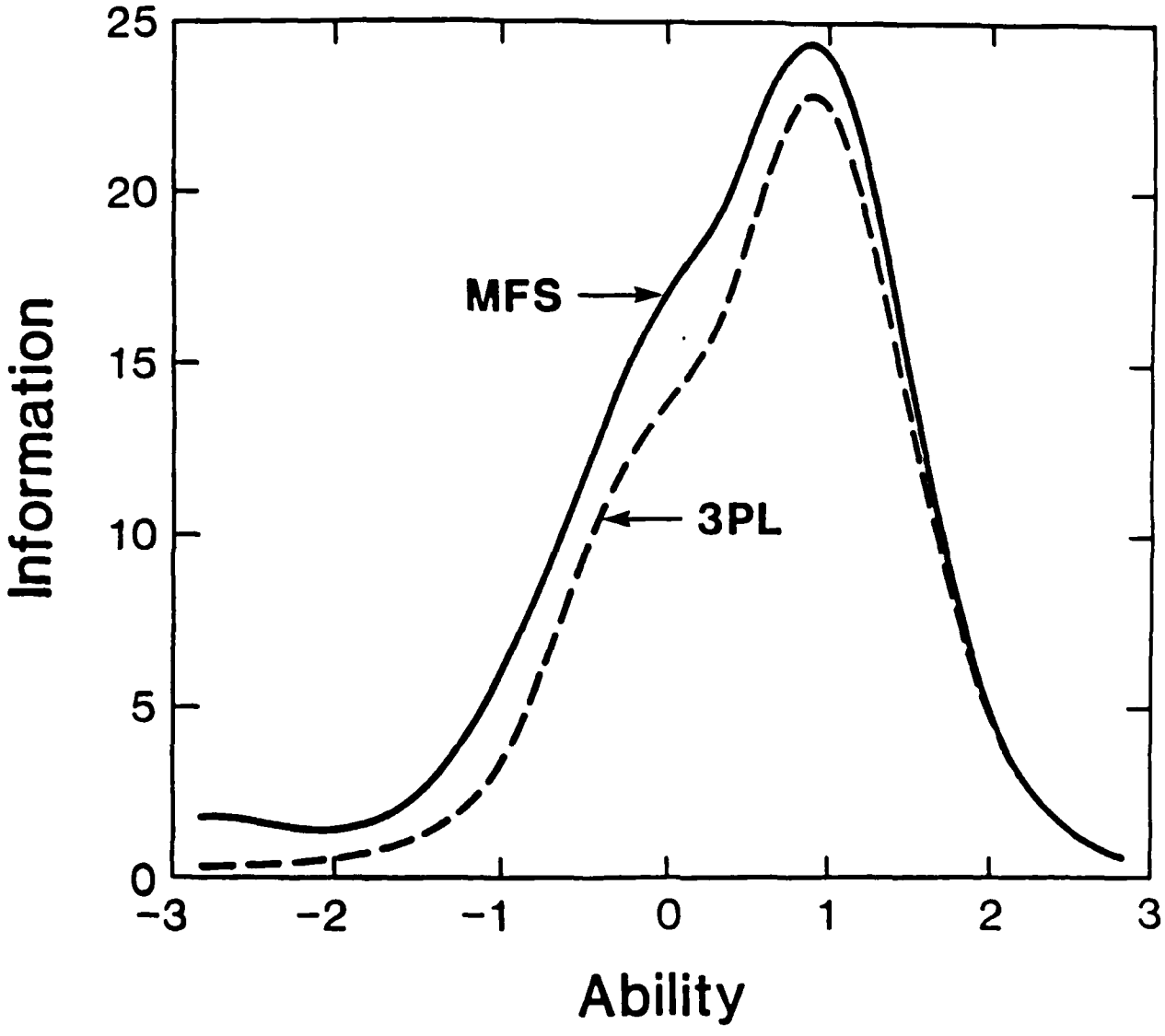


Figure 5. Information functions for dichotomous and polychotomous scorings of the Arithmetic Reasoning subtest.

aberrant response vectors at various false alarm rates were determined for each appropriateness index and each form of aberrance.

Appropriateness Indices

This section lists the appropriateness indices that are evaluated. Technical details about the indices are given in Chapter 2.

Polychotomous model optimal index (LR_p). Denote the polychotomously scored response vector by **v**. The polychotomous model optimal index studied here is

$$LR_p = \underline{P}_{Aberrant}(\mathbf{v}) / \underline{P}_{Normal}(\mathbf{v}),$$

where the probabilities are computed using three-parameter logistic ICCs to determine conditional probabilities of correct responses and MFS OCCs to determine conditional probabilities of incorrect responses.

Dichotomous model optimal index (LR_z). This index is identical to LR_p except that only the pattern of correct and incorrect responses **u** is used in its calculation. This class of indices, therefore, provides the highest rate of detection when the choice of incorrect option is ignored.

Dichotomous model optimal index computed using estimated item parameters (LR_z'). For optimal indices to be truly optimal, they must be computed using item parameters -- not item parameter estimates. In previous work (Levine & Drasgow, 1982), we found that the values of some appropriateness indices were almost unaffected when item parameter estimates were used in place of item parameters. In the present research, we also computed optimal indices for the three-parameter logistic model using estimated item parameters.

Dichotomous and polychotomous model standardized ℓ_0 (z, and z_p). In Chapter II, z, was discussed; z_p is the generalization of z, to the case of a polychotomous analysis of the item responses.

Fit statistics (F1 and F2). (Discussed in Chapter II.)

Caution indices (S, T2, and T4). (Discussed in Chapter II.)

Item-option variance (IOV). (Discussed in Chapter II.)

Likelihood function curvature statistics (JK and O/E). (Discussed in Chapter II.)

Deliberate failure key (DFK). The final index evaluated is the DFK developed by the Navy Personnel Research and Development Center (Swanson & Foley, 1982) to detect individuals who are deliberately attempting to obtain low scores. Although DFK was developed for the AFQT composite, we used the key for the AR subtest only.

Method

Data Sets. A test norming sample of 3,000 response vectors was created by sampling 3,000 numbers (θs) from the normal (0,1) distribution truncated to the [-5.0, 3.5] interval. A normal sample of 4,000 response vectors was also generated in this way. Then 2,000 aberrant response vectors were created in each of 16 conditions. These conditions resulted from varying three factors: the type of aberrance (spuriously high; spuriously low), the severity of aberrance (mild; moderate), and the distribution from which simulated abilities were sampled.

Eight of the aberrant samples contained spuriously high response vectors, and the remaining eight samples contained spuriously low response vectors. Spuriously high response patterns were created by first generating normal response vectors (using the AR three-parameter logistic ICCs to determine the probabilities of correct responses, and the AR COCCs to determine the probabilities of incorrect option selection) and then replacing either 17% (mild aberrance) or 33% (moderate aberrance) of the simulated responses (randomly sampled without replacement) with correct responses. Spuriously low response patterns were also created by first generating normal response vectors. Then 17% or 33% of the items were randomly selected without replacement and the responses to these items replaced with random responses (i.e., a response was replaced by option A with probability .25, by option B with probability .25, ..., and by option D with probability .25).

The third variable manipulated was the ability level of the aberrant sample. Abilities for the spuriously high samples were sampled from four parts of the normal (0,1) distribution truncated to [-5.0, 3.5]: very low (0th through 9th percentiles), low (10th through 30th percentiles), low average (31st through 48th percentiles), and high average (49th to 64th percentiles). In all cases, percentiles were determined after the truncation. Abilities were sampled from four average to high ability strata for the spuriously low samples: low average (31st to 48th percentiles), high average (49th through 64th percentiles), high (65th through 92nd percentiles), and very high (93rd percentile and above).

Analysis. The analysis followed the procedure described in Chapter II. All the item and test statistics required to compute the practical appropriateness indices were computed using the test norming sample. LOGIST (Wood et al., 1976) was used to estimate three-parameter logistic item parameters and a Fortran program was written to compute the other quantities required.

The practical appropriateness indices and LR_i' were then computed for the response vectors in the normal and aberrant samples. Optimal indices were also computed for the normal sample for four aberrant conditions: 17% spuriously high, 33% spuriously high, 17% spuriously low, and 33% spuriously low. The 17% spuriously high optimal index was computed for the four samples with this form of aberrance, the 33% spuriously high optimal index was computed for the four samples with this form of aberrance, etc. The ICCs and COCCs used to generated the data were used to compute LR_p and LR_i'.

Results

The results for the spuriously high conditions are given in Tables 10 through 13. The results for the lowest ability group are shown in Table 10. In this table, it is evident that cheating on five randomly selected items was not very detectable: At a 2% false alarm rate, only 28% of the simulated cheaters were detected by the optimal LR_p index. The best of the practical indices, z, and F2, detected 18% and 20%, respectively. (The higher detection rate of IOV resulted because this index is poorly standardized; see Chapter II.) Cheating on 10 items (the 33% condition) was reasonably detectable. For example, LR_p detected 61% and LR_3 detected 54% at a 2% false alarm rate. At this false alarm rate, z_3, F2, and T4 detected 44%, 41%, and 50%, respectively.

The detection rates of the optimal indices showed a relatively small decline from Table 10 to Table 11. At a 2% false alarm rate, LR_p, for example, declined from 28% to 26% for the 17% spuriously high treatment and from 61% to 53% for the 33% treatment. Most of the practical indices showed larger declines in detection rates. This trend continues in Table 12.

Finally, in Table 13, it is evident that simulated cheating on the AR subtest was almost undetectable for high average examinees. In contrast, Drasgow et al., (1985) found moderate detection rates for simulated cheaters with comparable abilities for the SAT-V. A significant difference between the two tests lies in the frequency (and relative frequency) of difficult ($\underline{\hat{b}}_i > 1.0$), discriminating ($\underline{\hat{a}} > 1.0$) items with low lower asymptotes ($\underline{\hat{c}} \leq .10$). Seventeen of the 85 SAT-V items satisfied these three conditions. In contrast, none of the 30 AR items met these conditions and only three items had $\underline{\hat{b}}_i > 1.0$. In sum, high average examinees had a reasonably good chance of responding correctly to each AR item; so, correct responses obtained by cheating were not clearly aberrant.

The results for the spuriously low samples are given in Tables 14 through 17. In Table 14, it is evident that 33% spuriously low responding by simulated low average examinees was moderately detectable by LR_p (a 30% detection rate with 2% false alarms) but not by any of the other appropriateness indices. Higher detection rates were obtained for simulated high average examinees (shown in Table 15). Again, LR_p performed substantially better than any other index. High rates of detection of simulated high and very high examinees are shown in Tables 16 and 17. LR_p was clearly the best index, with detection rates of 72% and 81% for a 2% false alarm rate in the 33% spuriously low treatment.

Table 10. Selected ROC Points for Spuriously High Response Patterns Generated from the 0-9% Ability Range

False alarm rate	Proportion detected by													
	LR_p	LR_z	LR_z'	z_p	z_z	F1	F2	S	T2	T4	IOV	JK	O/E	DFK
17% Spuriously High Treatment														
.001	04	04	01	00	03	00	01	00	00	01	10	00	00	00
.005	11	12	11	03	06	00	08	00	04	04	16	02	02	00
.01	16	19	17	05	12	02	13	03	07	06	23	03	04	03
.02	28	29	26	08	18	04	20	12	13	11	37	06	07	03
.03	34	33	30	11	25	07	24	18	18	14	45	09	09	12
.04	38	37	34	13	29	10	28	24	22	18	52	13	12	12
.05	43	40	38	15	33	15	32	27	26	22	57	15	14	12
.07	48	45	44	19	41	24	40	37	32	26	64	22	19	12
.10	52	50	49	26	51	36	50	49	42	33	71	29	25	28
33% Spuriously High Treatment														
.001	23	24	17	02	10	00	04	00	06	12	27	00	00	00
.005	40	33	27	07	25	00	15	00	28	27	37	00	04	00
.01	45	45	43	12	30	01	27	06	37	34	49	00	09	01
.02	61	54	52	17	44	05	41	17	50	46	66	01	17	01
.03	67	59	58	22	50	16	47	24	60	52	73	02	24	01
.04	71	64	63	25	56	23	55	32	65	57	80	03	37	01
.05	72	67	66	31	62	30	59	37	69	61	83	03	37	01
.07	77	71	70	37	66	42	68	47	74	67	84	07	47	01
.10	81	75	75	46	75	57	76	60	81	73	92	19	57	17

Table 11. Selected ROC Points for Spuriously High Response Patterns Generated from the 10-30% Ability Range

False alarm rate	Proportion detected by													
	LR_p	LR_3	LR_3'	z_p	z_3	F1	F2	S	T2	T4	IOV	JK	O/E	DFK
17% Spuriously High Treatment														
.001	02	01	00	00	02	00	00	00	00	01	04	00	00	00
.005	09	07	07	01	05	00	03	00	05	04	07	00	01	00
.01	14	14	14	04	09	00	06	00	07	07	13	00	03	00
.02	26	25	22	06	14	01	11	04	14	12	24	01	05	00
.03	31	29	29	08	19	03	14	06	20	16	31	02	07	03
.04	34	33	33	10	23	06	18	10	24	20	39	02	10	03
.05	40	36	37	12	27	09	21	12	27	23	43	03	14	03
.07	46	43	43	16	34	14	27	18	33	28	51	06	20	03
.10	52	50	51	23	43	24	37	28	42	34	61	14	27	12
33% Spuriously High Treatment														
.001	16	16	13	00	04	00	00	00	03	09	10	00	01	00
.005	31	27	23	03	14	00	07	00	20	23	17	00	06	00
.01	37	40	39	05	20	00	15	01	28	29	26	00	10	00
.02	53	50	50	08	30	03	27	06	42	41	40	00	20	00
.03	61	56	57	12	37	08	34	10	51	47	47	00	27	01
.04	65	63	62	14	42	12	42	16	58	53	56	00	34	01
.05	68	66	65	19	49	17	46	20	62	58	61	00	40	01
.07	73	70	70	25	54	28	56	29	67	63	68	05	51	01
.10	78	74	75	33	64	44	67	41	74	70	75	18	60	06

Table 12. Selected ROC Points for Spuriously High Response Patterns Generated from the 31-48% Ability Range

False alarm rate	Proportion detected by													
	LR_p	LR_z	LR_z'	z_p	z_z	F1	F2	S	T2	T4	IOV	JK	O/E	DFK
17% Spuriously High Treatment														
.001	00	00	00	00	01	00	00	00	00	01	00	00	00	00
.005	03	03	04	00	03	00	01	00	04	04	01	00	00	00
.01	06	07	08	02	06	00	02	00	06	06	04	00	01	00
.02	15	15	14	03	09	00	05	00	12	12	08	00	05	00
.03	20	19	19	05	14	03	07	02	17	15	12	00	08	00
.04	24	23	24	06	17	06	10	03	21	18	17	00	10	00
.05	29	26	28	07	20	07	13	04	23	22	20	00	13	00
.07	36	34	35	10	25	12	18	07	29	26	26	01	20	00
.10	43	42	43	15	33	18	26	12	36	32	35	07	29	06
33% Spuriously High Treatment														
.001	06	10	07	00	02	00	00	00	02	06	01	00	01	00
.005	17	16	14	01	07	00	03	00	12	16	03	00	05	00
.01	22	27	26	02	10	00	08	00	18	22	05	00	08	00
.02	39	36	37	04	17	04	16	02	27	32	11	00	17	00
.03	48	43	45	05	22	08	21	05	36	38	15	00	23	00
.04	53	51	49	07	27	12	27	07	41	43	16	00	29	00
.05	56	55	54	09	33	16	31	09	45	47	21	00	34	00
.07	63	61	61	13	37	23	40	14	50	53	25	07	44	00
.10	71	67	68	20	46	36	51	22	59	60	31	19	53	03

Table 13. Selected ROC Points for Spuriously High Response Patterns Generated from the 49-64% Ability Range

False alarm rate	Proportion detected by													
	LR_p	LR_z	LR_z'	z_p	z_z	F1	F2	S	T2	T4	IOV	JK	O/E	DFK
17% Spuriously High Treatment														
.001	00	00	00	00	00	00	00	00	00	00	00	00	00	00
.005	00	00	01	00	01	00	00	00	02	03	00	00	00	00
.01	02	01	03	00	03	01	01	00	04	04	00	00	00	00
.02	07	06	07	01	05	01	03	00	07	08	01	00	00	00
.03	11	09	11	01	08	04	04	00	11	11	02	00	06	00
.04	14	13	14	02	10	06	07	01	14	14	03	00	09	00
.05	18	16	17	03	13	08	08	01	16	17	04	00	12	00
.07	25	23	24	06	17	11	13	03	20	21	07	01	17	00
.10	33	30	34	09	23	16	19	05	26	27	11	07	24	03
33% Spuriously High Treatment														
.001	01	02	01	00	00	00	00	00	00	02	00	00	00	00
.005	05	04	03	00	03	01	01	00	05	07	00	00	01	00
.01	08	10	11	00	04	02	04	00	07	10	00	00	02	00
.02	19	16	18	01	07	07	08	01	12	17	01	00	06	00
.03	28	23	25	02	10	11	11	02	16	20	01	00	08	00
.04	34	32	32	03	12	14	15	03	20	25	03	00	11	00
.05	37	37	36	05	16	17	17	04	23	29	04	00	14	00
.07	48	45	46	08	19	23	23	07	28	35	05	03	20	00
.10	60	55	56	13	25	31	31	12	35	40	10	11	28	01

Table 14. Selected ROC Points for Spuriously Low Response Patterns Generated from the 31-48% Ability Range

False alarm rate	Proportion detected by													
	LR_p	LR_z	LR_z'	z_p	z_z	F1	F2	S	T2	T4	IOV	JK	O/E	DFK
17% Spuriously Low Treatment														
.001	01	00	00	00	00	00	00	00	00	00	00	00	00	00
.005	05	01	01	03	02	00	01	00	02	02	01	00	00	00
.01	09	03	03	05	04	01	02	00	03	03	03	01	01	01
.02	15	06	07	08	07	02	04	00	06	07	06	01	02	01
.03	18	10	12	12	10	04	05	01	09	09	08	02	03	06
.04	21	14	15	14	13	07	07	03	12	12	12	03	05	06
.05	24	17	18	15	15	10	09	04	14	14	14	05	07	06
.07	29	22	23	21	19	17	12	07	18	17	20	07	10	06
.10	35	28	28	27	26	25	17	11	23	22	26	12	14	20
33% Spuriously Low Treatment														
.001	07	01	01	01	02	00	00	00	00	01	02	00	00	00
.005	14	03	04	07	05	00	04	00	03	04	04	01	01	01
.01	22	08	09	12	10	02	07	00	05	07	07	02	01	01
.02	30	14	16	18	15	05	11	03	09	11	13	04	03	01
.03	36	20	22	23	20	09	13	06	14	15	18	07	04	16
.04	41	24	26	27	23	13	17	10	16	19	23	10	06	16
.05	45	29	30	31	26	17	19	11	19	22	27	13	07	16
.07	51	36	37	36	32	27	24	17	22	27	32	17	11	16
.10	59	44	44	44	38	36	31	25	29	33	41	24	16	37

Table 15. Selected ROC Points for Spuriously Low Response Patterns Generated from the 49-64% Ability Range

False alarm rate	Proportion detected by													
	LR_p	LR_z	LR_z'	z_p	z_z	F1	F2	S	T2	T4	IOV	JK	O/E	DFK
17% Spuriously Low Treatment														
.001	07	00	00	00	01	00	00	00	00	01	00	00	00	00
.005	16	04	04	04	03	00	00	00	03	03	01	00	00	00
.01	20	07	07	07	06	00	02	00	05	05	02	00	01	00
.02	26	14	13	11	09	03	04	00	10	08	06	00	02	00
.03	28	19	18	14	14	08	06	01	14	11	09	00	03	01
.04	31	24	21	16	16	13	08	02	17	14	13	00	05	04
.05	35	27	24	20	19	17	09	02	21	17	15	01	07	04
.07	39	31	29	25	24	24	13	05	25	21	20	01	13	04
.10	44	38	34	33	32	30	20	19	31	27	28	06	18	17
33% Spuriously Low Treatment														
.001	16	02	04	01	04	00	00	00	01	03	02	00	00	00
.005	25	07	11	12	09	00	03	00	08	08	05	00	00	00
.01	33	13	20	18	15	01	07	00	12	11	08	00	00	04
.02	40	19	27	26	21	06	11	02	19	18	17	01	00	04
.03	46	20	34	32	27	14	14	05	24	22	23	02	00	14
.04	50	29	38	36	30	22	18	08	27	27	29	03	07	14
.05	53	34	42	39	33	28	20	10	30	30	33	03	09	14
.07	59	40	47	46	39	38	26	16	34	35	38	05	15	14
.10	66	46	53	55	46	44	34	23	43	42	47	12	21	34

Table 16. Selected ROC Points for Spuriously Low Response Patterns Generated from the 65-92% Ability Range

False alarm rate	Proportion detected by													
	LR_p	LR_z	LR_z'	z_p	z_z	F1	F2	S	T2	T4	IOV	JK	O/E	DFK
17% Spuriously Low Treatment														
.001	20	05	04	00	02	00	00	00	01	02	00	00	00	00
.005	30	17	16	05	07	03	02	00	08	07	00	00	00	00
.01	34	24	22	08	13	08	06	01	12	10	02	00	00	00
.02	41	30	30	14	20	20	12	04	19	17	04	00	00	00
.03	43	34	33	19	26	28	15	06	24	20	05	00	06	01
.04	46	38	36	23	28	32	20	10	28	24	08	00	09	02
.05	49	40	39	26	31	36	22	12	31	27	10	00	12	02
.07	52	43	43	33	37	41	28	17	35	32	14	03	18	02
.10	56	49	49	43	45	46	38	22	43	38	20	10	24	09
33% Spuriously Low Treatment														
.001	38	14	17	03	15	00	01	00	08	12	06	00	00	00
.005	48	24	28	20	24	02	11	00	26	25	10	00	00	00
.01	55	34	38	29	36	08	19	04	33	31	15	00	08	02
.02	62	41	44	38	45	24	30	11	43	42	27	00	17	02
.03	65	47	50	44	51	36	37	15	51	46	33	00	22	09
.04	68	50	52	49	55	43	43	19	55	51	40	00	28	09
.05	71	54	55	53	59	49	46	23	58	54	43	00	32	09
.07	74	54	60	61	64	57	53	31	62	59	50	07	42	09
.10	78	64	65	69	71	64	61	41	69	65	58	22	50	27

Table 17. Selected ROC Points for Spuriously Low Response Patterns Generated from the 93-100% Ability Range

False alarm rate	Proportion detected by													
	LR_p	LR_z	LR_l	z_p	z_z	F1	F2	S	T2	T4	IOV	JK	O/E	DFK
17% Spuriously Low Treatment														
.001	45	22	22	04	04	11	01	00	02	03	00	00	00	00
.005	55	42	40	13	11	27	09	09	15	11	00	00	00	00
.01	60	49	46	18	20	43	18	22	21	18	00	00	00	00
.02	67	54	53	26	29	55	30	35	33	29	01	00	00	00
.03	69	58	56	32	37	60	35	41	41	35	01	00	00	00
.04	71	60	58	37	41	63	41	48	47	41	02	00	01	00
.05	72	62	60	40	46	66	45	51	51	47	03	00	01	00
.07	74	65	62	48	54	71	53	58	56	53	04	02	03	00
.10	77	68	66	58	63	75	63	65	64	62	06	11	06	04
33% Spuriously Low Treatment														
.001	64	42	40	04	32	02	06	00	20	33	09	00	02	00
.005	72	53	51	27	49	17	32	08	51	52	13	00	08	00
.01	76	61	59	39	62	36	46	21	59	60	20	00	13	00
.02	81	67	64	51	71	59	61	39	69	70	30	00	22	00
.03	83	70	68	59	77	69	67	48	74	74	35	00	27	05
.04	85	72	70	64	79	74	72	55	80	78	40	00	33	05
.05	86	74	73	68	82	78	75	59	83	81	44	00	38	05
.07	87	77	75	75	86	84	80	68	86	84	51	21	48	05
.10	90	79	77	82	90	87	86	76	90	87	58	41	57	21

Discussion

In this chapter, we described Levine's (1985a, 1985b) theory of psychological measurement. It was used to estimate COCCs for a sample of 2,891 examinees who responded to the AR subtest. Good to excellent fits were obtained when the estimated COCCs were compared to empirical proportions computed from the responses of a larger sample of 11,914 examinees. A simulation data set was also used to investigate COCC estimates. Very accurate estimates were obtained for ability ranges having sufficient numbers of examinees who responded incorrectly.

The test information function of the polychotomous model was found to be moderately larger than the three-parameter logistic information function for low to moderately high ability levels. Since there is information in incorrect options, it seems prudent to use it if items are expensive to write, if the number of items that can be administered is severely limited, or if very accurate ability estimates are required. Furthermore, we can now study systematically the differences between items with informative incorrect options and items with essentially noninformative incorrect options. It may be possible to identify different characteristics of these two types of items. Then item writers could explicitly attempt to write items with highly informative incorrect options and thus increase the information about ability provided by tests.

An Appropriateness Measurement simulation study was also conducted to compare the polychotomous model with a dichotomous submodel; namely, the three-parameter logistic. Several important results were obtained. First, for the spuriously low treatment that simulates atypical educations, misgridding answers to a portion of the test, unusual creativity, etc., we found that optimal three-parameter logistic appropriateness indices fell far short of their optimal polychotomous model counterparts. At some false alarm rates, the rates of detection of aberrant response vectors were more than 100% higher for the polychotomous optimal indices. Thus, Appropriateness Measurement constitutes one important practical testing problem where substantial gains are made by the use of a polychotomous item response model.

The results of the Appropriateness Measurement simulation study also showed that the practical polychotomous model index z_p was not a particularly good index: Its detection rates were not close to optimal for either spuriously high or spuriously low treatments. This result, in conjunction with the results described previously, points to the need to devise better polychotomous appropriateness indices that can be used in practical situations.

A third result obtained in the Appropriateness Measurement research reported in this chapter was that the z_3, F2, and T4 indices effectively detected aberrance in relation to three-parameter logistic optimal indices (but not polychotomous model optimal indices). Therefore, if one is satisfied with dichotomous scoring of item responses for some particular application, then z_3, F2, and T4 can be used with confidence to detect inappropriate test scores.

In sum, COCC estimates provide opportunities to improve testing in a variety of ways: ability estimation, the theory and practice of item writing, and Appropriateness Measurement. Applications in areas such as the evaluation of item and test bias and adaptive testing may also be fruitful. Consequently, we conclude that there is information in incorrect responses and that polychotomous item response models can make important contributions to psychological testing.

IV. MULTI-TEST EXTENSIONS OF PRACTICAL AND OPTIMAL APPROPRIATENESS INDICES

Introduction

This chapter describes methods for efficient detection of inappropriate test scores in situations where examinees complete several short tests. In particular, information about aberrance is pooled across tests that measure distinct traits. This approach seems valuable for test batteries such as the ASVAB, which contains a number of short power subtests.

Model-based approaches to the detection of aberrant response patterns have generally assumed that the latent trait space is unidimensional. For example, the three-parameter logistic model has been used by Levine and his colleagues (Drasgow & Levine, 1986; Drasgow et al., 1985; Levine & Drasgow, 1982; Levine & Rubin, 1979). Tatsuoka (Harnisch & Tatsuoka, 1983; Tatsuoka, 1984) has used the two- and three-parameter logistic models for her extended caution indices. Wright (1977) has tried to identify individuals who do not conform to another unidimensional model; namely, the Rasch model.

In Chapter II, we found that appropriateness indices can provide very high detection rates for long unidimensional tests. Detecting aberrant response patterns on shorter tests was shown to be a much more difficult task in Chapter III. What can be done to increase detection rates on short tests? The solution does not lie in better appropriateness indices for unidimensional tests, because no index computed from the item responses can provide higher detection rates than the optimal index used in Chapter III. This fact led us to devise methods for pooling information about aberrance across several short, unidimensional tests.

Another approach to detecting aberrant response patterns uses external information to predict test scores. The standardized residual (i.e., the standardized error of prediction) can then be used as an appropriateness index. For example, test scores not included in a selection composite can be used to predict the composite score. Persons who cheated on the tests included in the composite, but not on the other tests, would be expected to have large positive standardized residuals and therefore be identifiable. Similarly, scores from operational sections of a test can be used to predict scores on an experimental section in order to identify examinees who do not make a serious effort on the experimental section. These examinees would be expected to have large negative standardized residuals.

Little is known about the efficacy of the standardized residual approach to the identification of aberrant response patterns. In the second study described in this chapter, we evaluated this approach and compared it to model-based methods of Appropriateness Measurement.

The next section of this chapter describes multi-test extensions of six practical appropriateness indices, and then presents one means of approximating multi-test optimal indices. The approximation and multi-test practical indices were evaluated in two studies. The first used simulated ASVAB data so that all assumptions about the item responses (local independence, three-parameter logistic item characteristic curves, etc.) were correct. In the second study, an actual ASVAB data set was used so that the performances of the appropriateness indices could be evaluated under realistic conditions.

Multi-Test Extensions of Practical Appropriateness Indices

The basic assumption for our multi-test indices is that the test battery consists of several unidimensional tests. Let $\mathbf{U}_j = (\underline{U}_1, \ldots, \underline{U}_{n_j})$ denote the random vector of item responses for test $\underline{j}$, $\underline{j}=1, \ldots, \underline{m}$, let $\mathbf{u}_j = (\underline{u}_1, \ldots, \underline{u}_{n_j})$ denote a value of the random vector, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ denote a vector containing the abilities measured by each of the $\underline{m}$ tests. Then

$$\underline{P}(\mathbf{U}_1, \ldots, \mathbf{U}_m|\boldsymbol{\theta}) = \prod_{j=1}^{m} \underline{P}(\mathbf{U}_j|\boldsymbol{\theta})$$

$$= \prod_{j=1}^{m} \underline{P}(\mathbf{U}_j|\theta_j) ,$$

where both equalities result from local independence. This shows that the random vectors $\mathbf{U}_j$ are independent after conditioning on the individual abilities θ_j. Consequently,

$$\underline{P}(\underline{f}_1(\mathbf{U}_1), \ldots, \underline{f}_m(\mathbf{U}_m)|\boldsymbol{\theta}) = \prod_{j=1}^{m} \underline{P}(\underline{f}_j(\mathbf{U}_j)|\theta_j) , \qquad (23)$$

for arbitrary functions $\underline{f}_j$ (see Chung, 1974, p. 51), which means that functions of the item response are also conditionally independent.

Standardized ℓ_o. The significance of Equation 23 for developing multi-test extensions of appropriateness indices will be illustrated with the standardized ℓ_o indices. Let

$$\ell_o = \log \underline{P}(\mathbf{U}_1 = \mathbf{u}_1, \ldots, \mathbf{U}_m = \mathbf{u}_m|\boldsymbol{\theta})$$

$$= \sum_{j=1}^{m} \log \underline{P}(U_j = u_j|\theta_j)$$

$$= \sum_{j=1}^{m} \ell_o^{(j)} ,$$

where

$$\ell_o^{(j)} = \log \underline{P}(U_j = u_j|\theta_j) .$$

Then

$$E(\ell_o) = \sum_{j=1}^{m} E[\ell_o^{(j)}]$$

and by Equation 23

$$Var(\ell_o) = \sum_{j=1}^{m} Var[\ell_o^{(j)}] .$$

Hence, ℓ_o can be standardized by

$$\underline{z} = \frac{\ell_o - E(\ell_o)}{[Var(\ell_o)]^{1/2}} . \qquad (24)$$

Expressions for $E[\ell_o^{(j)}]$ and $Var[\ell_o^{(j)}]$ were given by Drasgow et al. (1985) for dichotomously and polychotomously scored item responses. We shall denote the standardized ℓ_o index by z, when the three-parameter logistic model is used. The index is denoted z_p when it is based on a polychotomous model.

In practice, the θ_j are not known. We have used maximum likelihood estimates $\hat{\theta}_j$ in place of the θ_j in our past research with apparent success (see Drasgow et al., 1985, Figures 3 and 4). Of course other approaches to estimation could be used. In fact, the well-known bias of maximum likelihood estimates suggests that perhaps alternative estimation methods should be explored.

Standardized extended caution indices. Let $T2^{(j)}$ and $T4^{(j)}$ denote Tatsuoka's (1984) second and fourth extended caution indices computed for the jth test. Tatsuoka found that $E[T2^{(j)}|\theta_j] = 0$ and provided expressions for $E[T2^{(j)}|\theta_j]$ and the conditional variances of $T2^{(j)}$ and $T4^{(j)}$. The standardized multi-test extensions of the two appropriateness indices are then

$$T2 = \frac{\Sigma\, T2^{(j)}}{[\Sigma\, \mathrm{Var}(T2^{(j)}|\theta_j)]^{1/2}} \tag{25}$$

and

$$T4 = \frac{\Sigma\, [T4^{(j)} - E(T4^{(j)}|\theta_j)]}{[\Sigma\, \mathrm{Var}(T4^{(j)}|\theta_j)]^{1/2}} . \tag{26}$$

Again, it is necessary to substitute estimates for the θ_j in Equations 25 and 26.

Fit statistics. The squared standardized residual fit statistic described by Wright (1977) involves an item-by-item standardization of the dichotomously scored item responses. Let $\underline{u}_{ij}$ equal 1 or 0 depending upon whether the examinee's response to item $\underline{i}$ on test $\underline{j}$ is correct or incorrect, let $\underline{P}_{ij}(\theta_j)$ equal the probability of a correct response to this item among examinees with ability θ_j, and let $\underline{Q}_{ij}(\theta_j) = 1 - \underline{P}_{ij}(\theta_j)$. Then a multi-test extension of Wright's statistic is

$$F1 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \{[\underline{u}_{ij} - \underline{P}_{ij}(\theta_j)]^2 / \underline{P}_{ij}(\theta)\underline{Q}_{ij}(\theta_j)\} . \tag{27}$$

The second fit statistic that we investigated was described by Rudner (1983). In our notation, this statistic is

$$F2^{(j)} = \underline{R}_j / \underline{V}_j ,$$

where

$$\underline{R}_j = \sum_{i=1}^{n_j} [\underline{u}_{ij} - \underline{P}_{ij}(\theta_j)]^2$$

and

$$\underline{V}_j = \sum_{i=1}^{n_j} \underline{P}_{ij}(\theta)\underline{Q}_{ij}(\theta) .$$

An extension to the multi-test case is

$$F2 = \sum_{j=1}^{m} \underline{R}_j \Big/ \sum_{j=1}^{m} \underline{V}_j . \tag{28}$$

Approximations to Optimal Appropriateness Indices

Unidimensional Tests. Levine and Drasgow (1984) showed that the most powerful appropriateness index for a given form of aberrance on a unidimensional test is the likelihood ratio statistic LR given in Equation 2. In our past research, we have evaluated the integrals in $\underline{P}_{Normal}(\mathbf{u})$ and $\underline{P}_{Aberrant}(\mathbf{u})$ by Simpson's rule, and used about 20 values of θ to give the likelihood ratio LR adequate accuracy. Although these numerical integrations are not particularly burdensome for a modern computer, generalizations to multi-test optimal indices would require excessive computations to evaluate multidimensional integrals. For this reason, we are led to seek a way to evaluate the integrals that will have a more convenient multi-test generalization.

Under general conditions, it can be shown that likelihood functions asymptotically (with the number $\underline{n}$ of items) have the shape of normal densities. Consequently, for long tests

$$\log \underline{P}_{Normal}(\mathbf{u}|\theta) \doteq a\theta^2 + b\theta + c \ . \qquad (29)$$

Throughout this chapter, we shall assume that the ability distribution $\underline{f}(\theta)$ is the standard normal, whence $\log[\underline{f}(\theta)]$ is a quadratic in θ. Therefore, both $\log[\underline{P}_{Normal}(\mathbf{u}|\theta)\cdot\underline{f}(\theta)]$ and $\log[\underline{P}_{Aberrant}(\mathbf{u}|\theta)\cdot\underline{f}(\theta)]$ should be approximately quadratic. The justification of this approximation lies in the high degree of agreement in Equation 29 and the high rates of detection of aberrant response patterns obtained in the present research. The computational details needed to reproduce our algorithm and replicate our results follow.

If

$$\log [\underline{P}_{Normal}(\mathbf{u}|\theta)\cdot\underline{f}(\theta)] \doteq \underline{a}\theta^2 + \underline{b}\theta + \underline{c} \qquad (30)$$

for $\underline{a} < 0$, then

$$\int \underline{P}_{Normal}(\mathbf{u}|\theta)\cdot\underline{f}(\underline{\theta})d\theta \doteq \int e^{(a\theta^2+b\theta+c)}d\theta$$

$$= e^c e^{b^2/2k^2} \int e^{-(\theta-b/k^2)^2/[2(1/k^2)]}d\theta$$

$$= \sqrt{-\pi/a}\, e^c e^{b^2/(-4a)} \ ,$$

where $k = \sqrt{-2a}$ and the last equality results from recognizing that the integrand in the previous equation is proportional to a normal density.

In order for this approximation to be accurate, the quadratic must fit well near the maximum of $\underline{y}(\theta) = \log [\underline{P}_{Normal}(\mathbf{u}|\theta)\cdot\underline{f}(\theta)]$. We used the following iterative procedure to obtain the quadratic. It begins by

evaluating $\underline{y}$ at five points: θ° = the maximum likelihood estimate $\hat{\theta}$ of θ; $\theta^\circ \pm .3$; and $\theta^\circ \pm .6$. Then a diagonal weight matrix is created with non-zero elements $\exp[\underline{y}(\theta)-\underline{y}(\theta^\circ)]$ corresponding to the five θ values. These weights are restricted to the interval [0.00001, 10.0] for computational reasons. Then the method of weighted least squares is used to obtain the initial coefficients $(a^\circ, b^\circ, c^\circ)$ of the quadratic.

The maximum of the fitted quadratic is $\theta' = -b^\circ/2a^\circ$. If θ' is within .15 of θ°, the iterative procedure ends; otherwise, five new θ values are selected as θ',

$$\theta' \pm \sqrt{(a^\circ)^{-1}\log(2/3)} ,$$

and

$$\theta' \pm \sqrt{(\underline{a}^\circ)^{-1}\log(1/3)} .$$

Then the weights are recomputed, and weighted least squares is used to obtain (a', b', c'). This process continues until $|\theta^{i+1} - \theta^i| \leq .15$. (Stricter convergence requirements did not seem to improve the approximation in Equation 30.)

Two restrictions are imposed to ensure convergence:

i) $a^i \leq -.01$;

and

ii) $|\theta^{i+1} - \theta^i| \leq 1.6/i$.

Convergence is usually obtained in one or two iterations.

Plotted in Figure 6 are 98 of 100 pairs of likelihood ratios. The abcissa values are the likelihood ratios that resulted from using Simpson's rule to evaluate $\underline{P}_{\text{Normal}}(\mathbf{u})$ and $\underline{P}_{\text{Aberrant}}(\mathbf{u})$; the ordinate values resulted from the quadratic approximations. The response patterns were simulated normal examinees responding to a 30-item test, item characteristic curves were three-parameter logistic ogives, ability was distributed as standard normal, and the form of aberrance was 15% spuriously low. The two pairs of points not plotted are (3.90, 3.91) and (5.07, 5.03).

In Figure 6, it is clear that the quadratic approximation was very accurate for likelihood ratios of less than 2.0. It was somewhat less accurate for larger values. In a variety of other tests, we found the approximation to be accurate for other aberrance hypotheses, for both simulated normal and simulated aberrant response patterns.

As a final check on the quadratic approximation, we determined hit rates for the 33% spuriously low condition using the item parameters from Chapter III. In this analysis, response vectors were generated from abilities in the 86 to 92 percentile range, and likelihoods were computed by Simpson's rule and

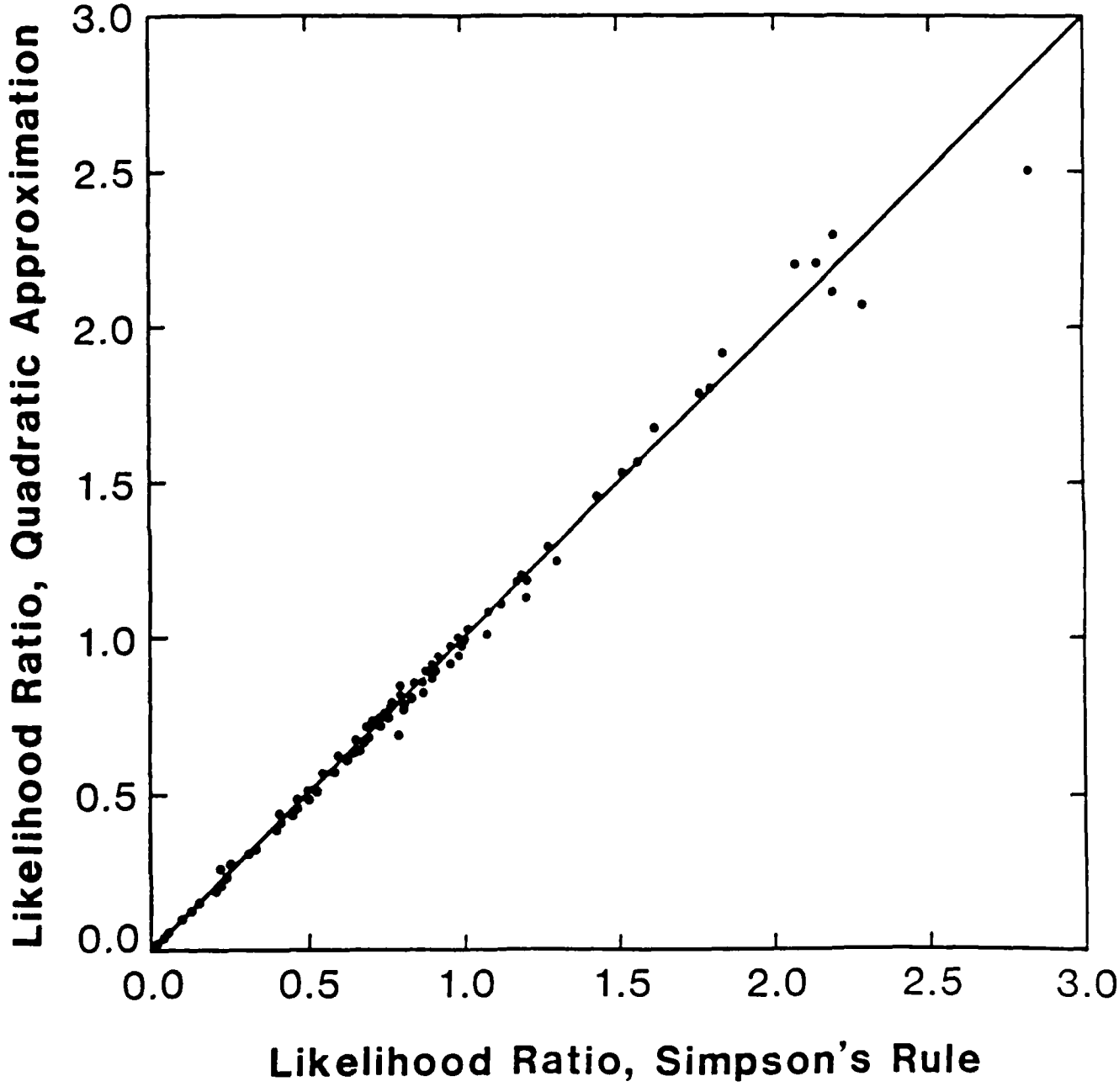


Figure 6. Likelihood ratios evaluated by Simpson's rule and the quadratic approximation for simulated normal response patterns.

by the quadratic approximation method. The detection rates at several false alarm rates are given below. LR_ℓ denotes the optimal index for the dichotomously scored item responses (ICCs were three-parameter logistic ogives), and LR_p denotes the optimal index for polychotomously scored item responses. It is clear that the quadratic approximation is sufficiently accurate for our purposes.

		False Alarm Rate				
Index	Method	.001	.01	.03	.05	.10
LR_p	Simpson	.53	.67	.75	.78	.84
LR_p	Quad. Approx.	.54	.66	.75	.79	.85
LR_ℓ	Simpson	.31	.53	.63	.68	.75
LR_ℓ	Quad. Approx.	.33	.51	.61	.66	.73

Two unidimensional tests. The likelihood that we must approximate is

$$\underline{F}^* = \int\int \underline{P}(\mathbf{U}_1 = \mathbf{u}_1|\theta_1)\ \underline{P}(\mathbf{U}_2 = \mathbf{u}_2|\theta_2)\ \phi_2(\boldsymbol{\theta};\mathbf{0},\boldsymbol{\Sigma})d\boldsymbol{\theta}\ , \qquad (31)$$

where $\underline{P}(\mathbf{U}_j = \mathbf{u}_j|\theta_j)$ is the likelihood of $\mathbf{u}_j$, $j = 1,2$, under either the normal or aberrant model, $\boldsymbol{\theta} = (\theta_1,\theta_2)'$, $\mathbf{0} = (0,0)'$,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

is the covariance matrix of the two traits, and ϕ_2 is the bivariate standard normal density,

$$\phi_2(\boldsymbol{\theta};\mathbf{0},\boldsymbol{\Sigma}) = (\det\boldsymbol{\Sigma})^{-1/2}(2\pi)^{-1}\ \exp[-\tfrac{1}{2}\,\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}]\ .$$

The final expression for the approximation and its derivation are given in Appendix C. The final expression depends only on the correlation ρ between θ_1 and θ_2, which is assumed to be known, and the coefficients $(\underline{a}_1, \underline{b}_1, \underline{c}_1)$ and $(\underline{a}_2, \underline{b}_2, \underline{c}_2)$ of the quadratic approximations that can be fitted to the likelihood functions of the two tests separately by the method described for a unidimensional test. Thus, we can fit quadratics to each separately by the method previously described and then easily compute the approximation to $\underline{F}^*$.

Study One: Simulated ASVAB Data

Purpose. How effective are the practical multi-test appropriateness indices relative to optimal multi-test appropriateness indices? What are the upper limits on the detectabilities of certain benchmark forms of aberrance when information from several short tests is combined?

In order for the optimal indices to be truly optimal, all assumptions used to specify the index must be true. For this reason, data were simulated

in Study One that perfectly satisfied all assumptions. In Study Two, an actual ASVAB data set was used so that we could evaluate the properties of the optimal and practical indices in realistic settings.

Data generation. The ASVAB AR subtest, the first of our two unidimensional tests, is a 30-item, four-option multiple-choice test. A sample of N = 2,978 examinees was taken from the NORC data set by selecting every fourth examinee (examinees 1, 5, 9, ...). The LOGIST (version 2B) computer program (Wood et al., 1976) was used to estimate three-parameter logistic ICCs. OCCs for the incorrect option (with omitted and not-reached treated as a single incorrect option) were estimated by means of Levine's (1985a; 1985b) MFS theory. A detailed description of these analyses was presented in Chapter III.

The 15-item Paragraph Comprehension subtest and the 35-item Word Knowledge subtest of the ASVAB were pooled to form our second unidimensional test. These two tests correlate .82 (Ree, Mullins, Mathews, & Massey, 1982), and their correlation corrected for attenuation is .96. Consequently, fitting unidimensional item response models to the pooled, 50-item Word Knowledge - Paragraph Comprehension (WKPC) subtest seemed justified.

As with the AR subtest, LOGIST was used to estimate ICCs, and MFS was used to estimate OCCs. Plots showing estimated curves and empirical proportions indicated good fits of both the ICCs and OCCs to the data.

The ICCs and OCCs estimated from the AR and WKPC subtests were used as the "true" ICCs and OCCs for the rest of Study One. As the first step in the simulation, a sample of 3,000 simulated response patterns was created and used as a test norming sample. The ICCs previously estimated were used to determine probabilities of correct responses, and the MFS OCCs were used to determine the probabilities of incorrect options. Abilities for the two tests were sampled from a bivariate standard normal distribution with the correlation parameter set equal to .8 (the correlations of WK and PC with AR are about .8 after correcting for unreliability; see Ree et al., 1982). Thus, for each simulated response pattern, a vector (θ_1, θ_2) was sampled from a bivariate standard normal with a correlation of .8; θ_1 and the AR ICCs and OCCs were used to simulate a polychotomously scored 30-item unidimensional test; and θ_2 and the WKPC ICCs and OCCs were used to simulate a polychotomously scored 50-item unidimensional test. The entire response vector of 80 items was taken as the data provided by one simulee.

The test norming sample was then used to determine the item and test statistics required to compute the multi-test practical appropriateness indices based on the three-parameter logistic model ($z_{,}$, T2, T4, F1, F2). This entailed two runs of LOGIST (one for the simulated AR and one for the simulated WKPC) and two runs of our own FORTRAN program.

A normal sample of 4,000 response vectors and 16 aberrant samples of 2,000 response vectors each were then created. The normal sample was generated exactly as was the test norming sample (except, of course, that different seeds were used for the random number generators). As in Chapters II and III, the aberrant samples resulted from varying three factors: the

type of aberrance (spuriously high; spuriously low), the severity of aberrance (mild; moderate), and the distribution from which simulated abilities were sampled.

Eight of the aberrant samples contained spuriously high response vectors, and the remaining eight samples contained spuriously low response vectors. Spuriously high response patterns were created replacing a given percentage $\underline{k}$ of simulated responses (randomly sampled without replacement) with correct responses for each of the two simulated unidimensional tests separately. Spuriously low response patterns were also created by applying the spuriously low manipulation to each of the two unidimensional tests separately. Mildly aberrant response patterns were generated by using $\underline{k}$ = 15% (i.e., 5 of 30 AR items and 8 of 50 WKPC items). Moderately aberrant response patterns were created using $\underline{k}$ = 30% (i.e., 9 of 30 AR items and 15 of 50 WKPC items).

The third variable manipulated was the ability level of the aberrant sample. A composite ability was computed for each examinee by the formula

$$\frac{\theta_1 + \theta_2}{[\mathrm{Var}(\theta_1 + \theta_2)]^{1/2}} = (\theta_1 + \theta_2)/1.9 \ .$$

Notice that the composite ability has a standard normal distribution. Composite abilities for the spuriously high samples were sampled from four parts of the standard normal distribution: very low (0th through 9th percentiles), low (10th through 30th percentiles), low average (31st through 48th percentiles), and high average (49th to 64th percentiles). Composite abilities were sampled from four average to high ability strata for the spuriously low samples: low average (31st to 48th percentiles), high average (49th through 64th percentiles), high (65th through 92nd percentiles), and very high (93rd percentile and above).

Analysis. The practical appropriateness indices were computed for the 4000 response vectors in the normal sample. The item and test statistics estimated from the test norming sample were used to compute all but one appropriateness index. The one exception was the standardized ℓ_o index computed from the polychotomously scored item responses, denoted z_p. It was computed using the true OCCs and ICCs. This allowed us to bypass estimation of OCCs from the test norming sample and provided a significant reduction in computing time. (Despite the advantage gained by being computed from true rather than estimated OCCs, it is shown below that z_p fell short of some other indices. Therefore, the advantage given to z_p was of little practical consequence.)

One non-IRT index was also computed: the Deliberate Failure Key (DFK), which was provided by the AFHRL.

Optimal appropriateness indices were computed (using the true OCCs and ICCs) for the normal sample for four aberrant conditions: 15% spuriously high, 30% spuriously high, 15% spuriously low, and 30% spuriously low. For each of these conditions two optimal appropriateness indices were computed.

The first, LR_p, is the optimal index for polychotomous scoring of the item responses. The second index, $LR_,$, results from using only the information in the dichotomously scored item responses. Thus, $LR_,$ is based on a submodel for the polychotomous data in which all the incorrect responses are grouped together.

The practical appropriateness indices were computed for each of the 16 aberrant samples. In addition, the three-parameter logistic and polychotomous model 15% spuriously high optimal indices were computed for the four samples with this form of aberrance, the 30% spuriously high optimal indices were computed for the four samples with this form of aberrance, etc.

Results. The results for the spuriously high conditions are given in Tables 18 through 21, and results for the spuriously low conditions are given in Tables 22 through 25. These tables show that the multi-test extensions provide sizable gains in detection rates. Table 18, which presents the results for the lowest ability range, illustrates this point. At a 1% false alarm rate for the 15% spuriously high condition, the polychotomous optimal index LR_p detected 22% of the aberrant response patterns if only the AR item responses were used, 37% from the WKPC item responses, and 55% from the combined 80 items. In Chapter II, we obtained a 50% detection rate under these conditions (15% spuriously high, 0 to 9th percentile ability range) for an 85-item unidimensional test. In fact, our polychotomous model, multi-test optimal index provided detection rates that are very similar to the rates obtained in Chapter II: At false alarm rates of 3%, 5% and 10%, our hit rates were 67%, 72%, and 78% for the 15% spuriously high treatment, respectively; the hit rates in Chapter II were 64%, 70% and 77%. For the 30% spuriously high treatment at false alarm rates of 1%, 3%, 5%, and 10%, the hit rates were 88%, 92%, 94% and 95%, respectively; the hit rates in Chapter II were 93%, 95%, 97% and 98%.

Comparisons of Tables 18 through 25 with our earlier results reveal that the polychotomous model, multi-test optimal indices provide detection rates that are generally similar to the rates provided by the polychotomous model optimal indices for the long unidimensional test. The differences that occur seem to be more due to the differences in the characteristics of the item pools (the items in the earlier study tended to be more difficult than the items used here) than to the dimensionality of the latent trait space (i.e., use of the multi-test extensions).

The hit rates for the multi-test practical appropriateness indices are less similar to the hit rates of practical indices on long unidimensional tests. The differences are particularly obvious for the spuriously high conditions. Perhaps the best way to illustrate the differences is to compare the detection rate of the best practical index to the detection rate of the optimal index. At a 1% false positive rate for the 30% spuriously high treatment in Table 18, this ratio equals .75 for z, divided by .88 for LR_p; namely, .75/.88 = .85. The corresponding ratio was .98 in Chapter II (.91 for T2 divided by .93 for LR_p). For the next higher ability range (10th through 30th percentiles), the ratio is .58 in Table 19; the corresponding ratio from Chapter II is .91. Finally, the ratio for the low average ability range from Table 20 is .47, and the ratio from Chapter II is .73.

Table 18. Selected ROC Points for Spuriously High Response Patterns Generated from the 00-09% Ability Range

False alarm rate	Test	Proportion detected by								
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	DFK
15% Spuriously High Treatment										
.001	AR	06	03	00	02	00	00	01	01	
	WKPC	19	07	00	05	00	01	01	03	
	MT	26	15	01	12	00	01	04	04	00
.01	AR	22	20	04	13	02	12	06	07	
	WKPC	37	22	04	24	00	10	07	09	
	MT	55	37	07	36	00	18	15	14	01
.03	AR	38	31	09	25	04	24	19	16	
	WKPC	49	35	10	41	00	22	17	17	
	MT	67	48	14	56	03	37	23	25	04
.05	AR	46	39	14	33	13	32	26	21	
	WKPC	57	41	15	50	00	30	24	23	
	MT	72	53	19	65	07	49	39	32	08
.10	AR	55	50	25	50	35	49	42	33	
	WKPC	66	48	25	63	13	47	37	35	
	MT	78	62	28	76	40	66	56	47	26
30% Spuriously High Treatment										
.001	AR	29	21	00	12	00	00	10	06	
	WKPC	42	19	00	21	00	01	14	17	
	MT	74	44	02	44	00	04	34	31	00
.01	AR	52	42	07	37	01	24	28	27	
	WKPC	68	41	07	50	00	18	33	34	
	MT	88	69	13	75	00	44	60	56	00
.03	AR	66	57	17	52	10	42	48	41	
	WKPC	79	53	15	67	00	39	50	48	
	MT	92	77	25	86	00	67	77	71	00
.05	AR	72	64	25	62	26	52	58	50	
	WKPC	82	59	22	76	03	50	60	55	
	MT	94	80	33	90	21	79	85	78	01
.10	AR	79	71	39	76	52	69	73	65	
	WKPC	86	64	34	84	34	70	73	69	
	MT	95	84	47	95	67	89	91	88	08

Table 19. Selected ROC Points for Spuriously High Response Patterns Generated from the 10-30% Ability Range

False alarm rate	Test	Proportion detected by								
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	DFK
15% Spuriously High Treatment										
.001	AR	04	03	00	01	00	00	01	00	
	WKPC	05	02	00	01	00	00	01	01	
	MT	10	07	00	03	00	00	03	03	00
.01	AR	17	17	02	10	01	06	07	08	
	WKPC	18	11	01	08	00	01	07	07	
	MT	37	27	02	16	00	04	12	12	00
.03	AR	33	28	06	19	03	14	18	16	
	WKPC	31	24	05	17	00	05	13	12	
	MT	50	42	06	31	01	13	23	22	00
.05	AR	40	37	09	26	09	20	24	22	
	WKPC	40	30	08	25	01	09	18	16	
	MT	58	49	09	38	04	22	32	28	00
.10	AR	51	49	19	40	25	34	40	33	
	WKPC	51	40	15	37	11	21	31	26	
	MT	66	59	17	52	21	38	47	41	03
30% Spuriously High Treatment										
.001	AR	20	18	00	07	00	00	08	05	
	WKPC	16	08	00	04	00	00	08	07	
	MT	50	31	00	15	00	00	22	20	00
.01	AR	42	37	01	26	00	15	23	24	
	WKPC	45	28	02	17	00	03	22	20	
	MT	74	59	04	39	00	17	43	42	00
.03	AR	60	53	10	40	07	29	40	37	
	WKPC	60	42	06	31	01	13	35	32	
	MT	82	70	11	59	04	37	60	58	00
.05	AR	67	61	15	49	17	38	49	47	
	WKPC	66	49	10	40	07	22	43	38	
	MT	86	74	16	67	11	53	69	67	00
.10	AR	75	71	27	65	38	57	65	62	
	WKPC	72	56	19	53	24	40	57	52	
	MT	89	80	26	78	41	69	80	78	00

Table 20. Selected ROC Points for Spuriously High Response Patterns Generated from the 31-48% Ability Range

False alarm rate	Test	Proportion detected by								
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	DFK
15% Spuriously High Treatment										
.001	AR	02	01	00	01	00	00	01	00	
	WKPC	00	00	00	00	00	00	01	01	
	MT	02	02	00	00	00	00	03	02	00
.01	AR	09	11	00	07	00	03	06	06	
	WKPC	05	04	00	02	00	00	05	04	
	MT	21	17	01	06	00	02	10	10	00
.03	AR	24	21	03	14	02	08	16	14	
	WKPC	16	14	02	07	02	03	11	10	
	MT	35	31	03	16	03	08	20	20	00
.05	AR	34	30	05	19	07	13	21	21	
	WKPC	24	21	05	12	07	05	17	15	
	MT	45	40	06	22	06	14	27	27	00
.10	AR	46	45	13	33	18	26	36	35	
	WKPC	39	33	12	22	17	14	28	25	
	MT	59	51	12	34	19	27	39	40	00
30% Spuriously High Treatment										
.001	AR	11	09	00	03	00	00	04	03	
	WKPC	02	01	00	00	00	00	03	03	
	MT	21	11	00	02	00	00	09	10	00
.01	AR	30	27	01	15	01	07	15	18	
	WKPC	19	12	01	04	01	01	11	11	
	MT	53	42	01	14	01	07	23	25	00
.03	AR	48	43	05	26	06	18	29	32	
	WKPC	35	24	03	12	05	05	19	19	
	MT	69	58	05	28	07	18	36	39	00
.05	AR	57	53	08	35	13	27	38	41	
	WKPC	45	32	06	19	12	10	26	25	
	MT	75	64	08	36	14	29	46	48	00
.10	AR	70	66	17	50	31	45	52	54	
	WKPC	56	44	14	29	25	23	38	37	
	MT	80	72	16	51	32	47	62	64	00

Table 21. Selected ROC Points for Spuriously High Response Patterns Generated from the 49-64% Ability Range

False alarm rate	Test	Proportion detected by								
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	DFK
15% Spuriously High Treatment										
.001	AR	00	00	00	00	00	00	00	00	
	WKPC	00	00	00	00	00	00	00	00	
	MT	00	00	00	00	00	00	01	01	00
.01	AR	04	05	00	04	00	01	04	05	
	WKPC	01	01	00	00	01	00	03	03	
	MT	06	06	00	01	01	01	04	05	00
.03	AR	14	12	01	09	04	05	10	11	
	WKPC	05	05	02	03	05	02	08	08	
	MT	18	17	02	06	05	04	11	12	00
.05	AR	22	19	03	13	08	09	14	16	
	WKPC	11	11	04	06	10	04	12	11	
	MT	30	25	04	10	09	09	16	17	00
.10	AR	37	33	09	23	16	20	25	26	
	WKPC	24	22	09	13	19	12	20	20	
	MT	47	42	08	19	19	19	28	29	00
30% Spuriously High Treatment										
.001	AR	03	03	00	00	00	00	01	01	
	WKPC	00	00	00	00	00	00	01	01	
	MT	05	03	00	00	00	00	02	03	00
.01	AR	15	14	00	08	02	04	08	11	
	WKPC	06	05	00	01	02	00	05	05	
	MT	31	23	00	03	02	04	09	12	00
.03	AR	32	27	02	16	08	11	17	22	
	WKPC	18	12	02	05	08	03	10	10	
	MT	48	39	03	11	09	11	19	23	00
.05	AR	42	36	05	21	14	18	22	27	
	WKPC	27	22	04	08	14	06	15	15	
	MT	59	49	05	17	15	19	28	30	00
.10	AR	58	54	11	34	27	30	34	37	
	WKPC	41	34	10	14	24	15	25	27	
	MT	69	59	12	29	27	32	42	44	00

Table 22. Selected ROC Points for Spuriously Low Response Patterns Generated from the 31-48% Ability Range

False alarm rate	Test	Proportion detected by								
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	DFK
15% Spuriously Low Treatment										
.001	AR	05	00	00	01	00	00	00	01	
	WKPC	07	02	00	00	00	00	03	02	
	MT	20	02	01	01	00	00	01	01	00
.01	AR	13	04	07	06	01	03	04	04	
	WKPC	24	08	10	08	00	01	07	07	
	MT	32	08	16	12	00	02	06	09	00
.03	AR	22	10	14	11	05	07	09	11	
	WKP	37	23	22	19	02	05	13	13	
	MT	47	26	29	19	03	07	14	18	00
.05	AR	26	18	20	16	10	10	15	15	
	WKPC	45	32	31	27	07	12	18	19	
	MT	54	38	36	29	09	13	21	24	01
.10	AR	39	29	30	26	22	19	24	24	
	WKPC	56	46	43	37	23	20	29	29	
	MT	65	55	52	41	27	25	33	33	06
30% Spuriously Low Treatment										
.001	AR	06	00	00	02	00	00	02	01	
	WKPC	17	07	03	04	00	00	04	06	
	MT	38	16	08	08	00	00	04	07	00
.01	AR	21	08	11	11	02	06	06	08	
	WKPC	49	25	25	20	00	03	12	16	
	MT	62	33	36	26	00	06	13	20	00
.03	AR	38	24	22	19	09	12	15	16	
	WKPC	59	44	43	36	02	12	20	27	
	MT	73	52	53	41	06	17	25	31	02
.05	AR	46	32	30	24	19	18	18	22	
	WKPC	67	52	55	45	09	18	26	33	
	MT	80	63	64	49	15	27	33	41	05
.10	AR	59	47	44	36	35	29	29	34	
	WKPC	80	63	69	57	35	35	39	45	
	MT	88	75	77	60	43	42	45	54	18

Table 23. Selected ROC Points for Spuriously Low Response Patterns Generated from the 49-64% Ability Range

False alarm rate	Test	Proportion detected by								
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	DFK
15% Spuriously Low Treatment										
.001	AR	03	00	00	01	00	00	01	01	
	WKPC	18	07	00	01	00	00	06	05	
	MT	25	04	02	03	00	00	05	05	00
.01	AR	17	06	07	07	02	02	05	05	
	WKPC	44	25	11	12	01	02	16	14	
	MT	50	27	18	17	01	03	16	15	00
.03	AR	27	15	14	13	08	07	13	12	
	WKPC	55	41	26	28	13	10	26	22	
	MT	63	45	34	31	15	11	27	27	00
.05	AR	32	24	20	18	16	11	17	17	
	WKPC	60	47	38	37	26	17	32	29	
	MT	67	52	44	40	27	20	36	34	00
.10	AR	43	34	33	29	29	20	29	27	
	WKPC	67	57	53	50	49	33	44	41	
	MT	75	63	58	52	51	34	48	46	01
30% Spuriously Low Treatment										
.001	AR	11	01	01	03	00	00	03	01	
	WKPC	37	20	06	11	00	00	14	16	
	MT	60	34	16	16	00	00	14	16	00
.01	AR	32	14	15	15	02	07	10	10	
	WKPC	66	43	33	35	00	08	27	30	
	MT	78	52	50	41	00	12	30	35	00
.03	AR	46	29	29	24	13	13	20	20	
	WKPC	75	59	54	52	09	23	40	43	
	MT	85	68	69	59	15	27	47	50	00
.05	AR	53	37	37	30	26	19	27	27	
	WKPC	79	66	67	61	27	32	48	50	
	MT	89	74	77	66	32	40	55	58	01
.10	AR	64	48	53	42	40	31	39	39	
	WKPC	87	76	80	72	57	53	60	62	
	MT	93	83	86	76	64	59	67	69	05

Table 24. Selected ROC Points for Spuriously Low Response Patterns Generated from the 65-92% Ability Range

False alarm rate	Test	Proportion detected by								
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	DFK
15% Spuriously Low Treatment										
.001	AR	14	04	00	02	00	00	02	01	
	WKPC	42	27	03	05	00	00	18	14	
	MT	55	27	08	12	00	00	19	14	00
.01	AR	34	19	08	13	13	05	09	09	
	WKPC	66	49	22	25	16	09	35	30	
	MT	74	56	30	34	22	14	38	34	00
.03	AR	44	31	19	22	28	14	21	19	
	WKPC	73	61	42	45	43	25	50	42	
	MT	81	69	52	54	53	31	53	47	00
.05	AR	49	38	26	29	36	20	27	25	
	WKPC	75	65	54	55	58	35	58	49	
	MT	84	73	63	62	63	43	62	56	00
.10	AR	56	46	42	42	47	33	42	36	
	WKPC	80	72	69	69	74	56	68	63	
	MT	87	79	76	73	79	59	73	69	00
30% Spuriously Low Treatment										
.001	AR	28	10	01	11	00	00	11	07	
	WKPC	61	42	10	22	00	01	39	37	
	MT	81	63	29	41	00	02	48	43	00
.01	AR	48	31	22	30	12	16	26	24	
	WKPC	81	64	46	55	02	25	57	57	
	MT	89	76	67	69	07	39	66	66	00
.03	AR	61	45	38	42	31	28	40	38	
	WKPC	85	75	68	71	30	48	69	67	
	MT	92	84	82	82	48	60	78	77	00
.05	AR	67	50	47	48	44	37	46	45	
	WKPC	88	79	79	78	54	60	75	73	
	MT	94	87	88	86	65	72	83	83	00
.10	AR	74	60	63	61	57	52	60	56	
	WKPC	93	85	89	86	79	76	82	82	
	MT	96	91	94	92	85	84	90	89	01

Table 25. Selected ROC Points for Spuriously Low Response Patterns Generated from the 93-100% Ability Range

False alarm rate	Test	Proportion detected by								
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	DFK
15% Spuriously Low Treatment										
.001	AR	49	26	05	06	10	00	07	06	
	WKPC	66	46	07	05	02	00	39	24	
	MT	89	69	16	18	09	03	44	29	00
.01	AR	59	48	19	23	43	17	23	21	
	WKPC	80	62	38	35	40	18	57	48	
	MT	93	78	62	62	58	38	66	62	00
.03	AR	69	57	34	37	60	36	37	37	
	WKPC	85	74	57	59	70	40	70	63	
	MT	94	86	78	73	84	62	83	78	00
.05	AR	72	61	43	47	66	42	48	46	
	WKPC	86	77	71	69	80	56	78	71	
	MT	95	88	84	83	89	72	89	84	00
.10	AR	77	67	60	62	73	59	64	60	
	WKPC	89	82	83	80	89	67	87	81	
	MT	96	91	93	91	95	84	94	90	00
30% Spuriously Low Treatment										
.001	AR	56	34	03	30	02	00	29	22	
	WKPC	78	61	12	27	00	00	61	62	
	MT	96	89	45	70	00	07	82	78	00
.01	AR	75	59	33	55	46	41	57	53	
	WKPC	90	79	58	70	00	43	82	79	
	MT	98	93	85	92	39	77	94	93	00
.03	AR	82	70	53	70	68	60	69	68	
	WKPC	93	85	78	85	55	71	89	86	
	MT	98	95	94	96	84	91	97	96	00
.05	AR	85	74	63	76	77	69	75	74	
	WKPC	94	87	87	90	76	80	92	90	
	MT	98	96	96	98	92	95	98	97	00
.10	AR	89	80	80	86	85	81	85	83	
	WKPC	96	90	94	95	92	91	96	94	
	MT	99	97	98	99	97	98	99	99	00

Discussion. The comparisons of the detection rates of the multi-test practical indices to rates for LR_p show an important difference between unidimensional Appropriateness Measurement and multidimensional Appropriateness Measurement. Specifically, z_3, T2, and T4 efficiently detected spuriously high response patterns on the long unidimensional SAT-V. Tables 18 through 21 show that we did not replicate this finding with the short AR and WKPC tests: There are substantial differences in hit rates between practical and optimal multi-test appropriateness indices. This finding provides a motivation for seeking better practical appropriateness indices.

Study Two: Actual ASVAB Data

Purpose. Do the results obtained for simulated ASVAB data generalize to actual ASVAB data? In previous research (Drasgow et al., 1985; Levine & Drasgow, 1982), we found that unidimensional ℓ_o appropriateness indices provided similar rates of detection with real and simulated data. Will we obtain similar results for the multi-test extensions of the standardized ℓ_o index and the other appropriateness indices?

For an optimal appropriateness index to be truly optimal, ICCs (and OCCs if the analysis is polychotomous) must be known and must fit the data, tests assumed to be unidimensional must be truly unidimensional, the correlation between ability on test one and ability on test two must be known, and the ability density must be known. We violated all of these conditions in Study Two. To what extent will detection rates for optimal indices be degraded?

Data sets. The NORC sample provided the data base for Study Two. The test norming sample consisted of responses of the N = 2,978 NORC examinees analyzed in the first phase of Study One. The AR and WKPC ICCs and OCCs estimated from this sample were used for all analyses in Study Two. Also, the statistics needed for the T2 and T4 indices were obtained from this sample. Finally, a standardized residual (SR) measure was created by first regressing the total number-right score from the AR and WKPC subtests on the Math Knowledge (MK) and General Science (GS) subtests of the ASVAB,

$$\text{Predicted (AR + WKPC)} = \hat{B}_1 + \hat{B}_2\text{MK} + \hat{B}_3\text{GS}$$

$$= 7.98 + 1.20\text{MK} + 1.88\text{GS} ,$$

and then standardizing the residual

$$\text{AR + WKPC - Predicted (AR + WKPC)}$$

as described by Cook and Weisberg (1982). The correlation between MK and AR, after correcting for attenuation, is .88; the corrected correlation between GS and WK is .94; and the corrected correlation between GS and PC is .90 (Ree et al., 1982). Large positive values of SR were used to indicate spuriously high test scores, and large negative values of SR were taken to indicate spuriously low scores.

A normal sample of 2,716 response vectors was formed by selecting every fourth examinee (2, 6, 10, ...) from the NORC sample, and then deleting the data from the 262 examinees who failed to answer at least 77% of the items on both the AR and the WKPC subtests. The requirement that examinees answer at least 77% of the items is based on the Drasgow et al. (1985) conclusion that test scores of individuals who answer less than 77% of the test are very likely to be invalid measures of ability.

The remaining examinees from the NORC sample (examinees 3, 4, 7, 8, 11, 12,...) were used to form six more samples. These samples were created by first determining the frequency distribution of total score across both the AR and WKPC subtests (i.e., AR + WKPC); sorting into groups on the basis of the percentiles used for the AFQT Categories; and finally, removing examinees who answered fewer than 77% of the items on either the AR or WKPC subtests. Score ranges and sample sizes for the six groups were:

Sample	AR + WKPC Score Range	Sample Size
very high	74 to 80	494
high	59 to 73	1537
high average	50 to 58	941
low average	39 to 49	959
low	24 to 38	1155
very low	0 to 23	342

Aberrant samples were formed exactly as in Study One. Thus, the 15% and 30% spuriously high treatments were applied to the four lowest ability groups, and the 15% and 30% spuriously low treatments were applied to the four highest ability groups.

Analysis. Appropriateness indices were computed as in Study One, with the main exception that optimal indices were computed with ICCs and OCCs estimated from the test norming sample. The correlation between θ_1 and θ_2 was assumed to be .8, and the ability density was assumed to be the standard normal truncated to (-5.0, 3.5). Appropriateness indices were computed for the six samples stratified on ability, before the aberrance treatments as well as after each aberrance treatment.

Index standardization. Although each practical appropriateness index (except DFK) was standardized, the expressions for the conditional expectations and variances of the indices were obtained using the assumption that θ_1 and θ_2 were known. Of course, in practice, they are unknown; therefore, it is important to investigate the conditional distributions of the appropriateness indices for normal examinees.

The standardizations of the practical indices can be determined from Figure 7. This figure presents ROC curves for seven practical appropriateness

indices: z_p, z_3, F1, F2, T2, T4, and SR. Abcissa values in all cases were determined from the normal sample of 2,716 examinees. For the top row of the figure, ordinate values were based on the responses of the 342 examinees in the very low ability range prior to any aberrance manipulation (i.e., this sample was simply a normal, low ability group). Ordinate values for the middle row of the figure were based on the low average sample, and the bottom row was determined from the very high ability sample. Response patterns were presumably normal for these two samples as well (we had not applied any aberrance treatment). Only the lower left quarter of each ROC curve is shown, in order to conserve space and because we are primarily concerned with an index's standardization for low misclassification rates. Results for the other three ability ranges are not shown because they were consistent with the trends that are apparent in Figure 7.

In Figure 7, it is clear that z_p, SR, and F1 are not consistently well standardized; z_3 is reasonably well standardized across ability levels, although its performance for the highest ability level is somewhat disappointing; and F2 is fairly well standardized across ability levels. The most surprising results are the very accurate standardizations of the multi-test extensions of T2 and T4. Their standardizations were not very good for the long unidimensional test studied in Chapter II; here, their standardizations are excellent except, perhaps, for the highest ability group.

Detection of aberrant response patterns. Tables 26 through 33 present the detection rates for the multi-test appropriateness indices when they are applied to actual ASVAB data. Comparing the results for the spuriously high conditions for real data (Tables 26 through 29) to the results for simulation data (Tables 18 through 21) reveals generally similar detection rates. The detection rates for the polychotomous model optimal index LR_p tended to be moderately decreased for the actual ASVAB data, but detection rates for the dichotomous model appropriateness indices were relatively unchanged.

Of the practical appropriateness indices, z_3 is clearly the most effective for the lowest ability range. The T2 and T4 indices had detection rates comparable to z_3 in the 10% to 30% ability range and appear slightly superior for the low average and high average ability ranges. The other five practical appropriateness indices (z_p, F1, F2, SR, and DFK) all had detection rates far lower than z_3, T2, and T4.

Although the detection rates for the spuriously high conditions are similar across the simulated and real data sets, there is an important difference: Both the normal and the aberrant groups for the actual ASVAB data sets had generally larger index scores. For example, 1.6% of the 4,000 simulated normals from Study One had z_3 scores less than -2.0, and 11.4% had z_3 scores less than -1.0 . For the 2,716 NORC examinees taken as the normal group, the corresponding rates were 3.4% and 16.2%. This trend was also apparent for T2, T4, and the three-parameter logistic optimal index. For example, LR_3 had 4.2% and 12.9% of its values greater than 5 and 2, respectively, for the NORC normals, versus only 1.8% and 7.7% for the Study One simulated normals.

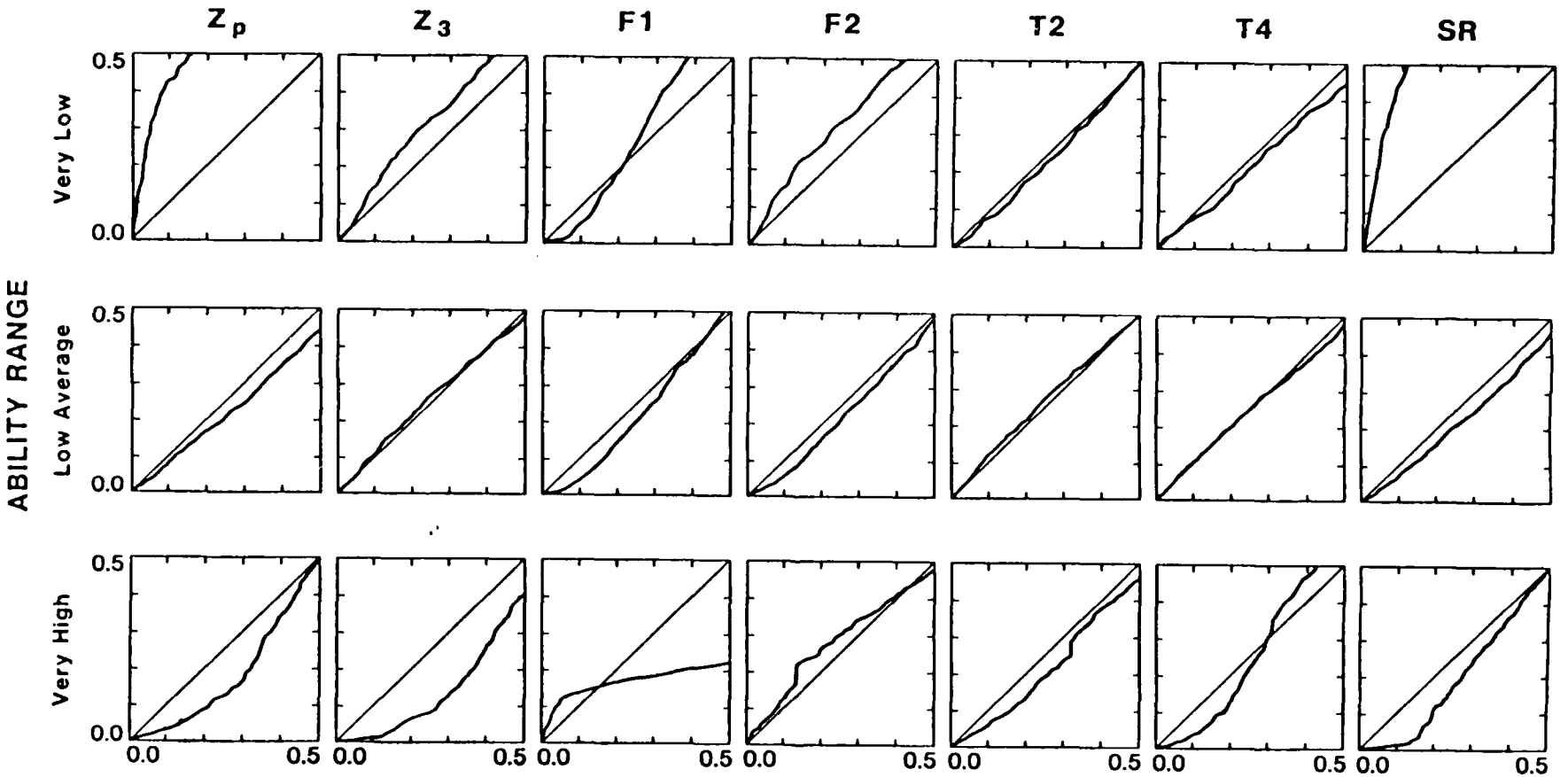


Figure 7. Standardizations of practical appropriateness indices.

Table 26. Selected ROC Points for Spuriously High Response Patterns Created from NORC Examinees in the 00-09% Ability Range

False alarm rate	Test	Proportion detected by									
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	SR	DFK
15% Spuriously High Treatment											
.001	AR	06	05	00	02	00	00	01	00		
	WKPC	06	04	00	00	00	04	00	00		
	MT	13	07	02	09	00	03	01	02	00	00
.01	AR	17	17	05	14	00	11	09	07		
	WKPC	24	21	09	26	00	17	09	10		
	MT	38	32	11	38	00	25	18	11	00	03
.03	AR	35	28	15	29	02	31	18	14		
	WKPC	36	36	22	43	00	32	18	22		
	MT	53	50	27	61	00	42	29	25	00	11
.05	AR	42	39	22	39	11	39	23	18		
	WKPC	39	41	32	54	02	37	25	29		
	MT	58	58	38	69	08	59	40	33	02	18
.10	AR	58	51	34	56	33	52	40	30		
	WKPC	49	52	47	69	23	55	41	46		
	MT	65	68	54	79	42	71	59	46	05	37
30% Spuriously High Treatment											
.001	AR	26	10	00	10	00	01	04	00		
	WKPC	25	08	00	16	00	05	12	15		
	MT	52	48	01	48	00	10	32	20	01	00
.01	AR	48	39	04	39	00	21	29	25		
	WKPC	51	42	08	60	00	21	41	41		
	MT	82	72	17	75	00	43	64	56	04	00
.03	AR	67	58	21	55	07	45	46	39		
	WKPC	62	58	28	73	00	46	55	60		
	MT	84	81	32	89	00	65	75	72	14	02
.05	AR	70	64	32	65	21	54	57	45		
	WKPC	64	64	37	79	04	51	61	69		
	MT	88	84	47	93	25	80	83	77	25	03
.10	AR	76	73	45	81	51	69	71	61		
	WKPC	70	70	52	89	42	72	77	82		
	MT	90	89	61	96	67	90	93	88	39	20

Table 27. Selected ROC Points for Spuriously High Response Patterns Created from NORC Examinees in the 10-30% Ability Range

False alarm rate	Test	Proportion detected by									
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	SR	DFK
15% Spuriously High Treatment											
.001	AR	02	03	00	01	00	00	00	00		
	WKPC	04	02	00	02	00	00	02	01		
	MT	06	03	00	05	00	01	02	03	00	00
.01	AR	15	14	01	09	00	06	05	05		
	WKPC	17	16	02	16	00	03	12	11		
	MT	28	26	03	19	00	06	18	12	01	00
.03	AR	30	29	07	21	01	17	17	13		
	WKPC	31	33	06	27	00	11	21	22		
	MT	47	45	07	38	00	14	27	24	03	00
.05	AR	41	39	12	30	06	24	22	17		
	WKPC	38	39	11	34	01	14	27	28		
	MT	55	54	12	46	03	26	36	31	06	01
.10	AR	54	52	20	45	19	36	38	29		
	WKPC	51	53	20	47	10	27	42	42		
	MT	66	65	24	60	18	44	54	47	16	07
30% Spuriously High Treatment											
.001	AR	18	11	00	07	00	00	04	01		
	WKPC	13	09	00	05	00	01	09	07		
	MT	33	35	00	17	00	01	18	10	02	00
.01	AR	43	35	01	26	00	09	23	21		
	WKPC	33	30	01	23	00	07	25	24		
	MT	64	58	03	38	00	15	46	40	09	00
.03	AR	60	55	10	43	03	31	43	41		
	WKPC	52	49	05	34	00	18	38	38		
	MT	77	71	09	63	00	34	59	60	23	00
.05	AR	68	63	16	54	11	41	51	47		
	WKPC	61	55	10	43	04	22	45	45		
	MT	82	79	15	70	10	50	70	69	33	00
.10	AR	75	73	29	71	32	59	67	62		
	WKPC	68	65	21	61	21	38	61	61		
	MT	88	86	30	82	36	70	82	80	50	02

Table 28. Selected ROC Points for Spuriously High Response Patterns Created from NORC Examinees in the 31-48% Ability Range

False alarm rate	Test	Proportion detected by									
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	SR	DFK
15% Spuriously High Treatment											
.001	AR	01	01	00	01	00	00	00	00		
	WKPC	00	00	00	00	00	00	01	00		
	MT	01	00	00	00	00	00	01	01	00	00
.01	AR	08	08	01	04	00	01	05	04		
	WKPC	03	02	00	03	00	01	04	04		
	MT	09	07	01	03	00	01	08	06	02	00
.03	AR	19	18	04	12	01	07	14	12		
	WKPC	12	11	01	06	01	03	09	11		
	MT	24	21	02	13	01	03	16	16	09	00
.05	AR	27	27	07	19	03	13	20	16		
	WKPC	19	16	02	08	02	04	14	15		
	MT	33	30	03	17	03	10	21	21	15	00
.10	AR	41	39	13	32	12	23	31	28		
	WKPC	33	31	08	18	10	10	24	24		
	MT	50	47	10	28	11	21	34	32	27	00
30% Spuriously High Treatment											
.001	AR	09	06	00	02	00	00	01	00		
	WKPC	01	01	00	00	00	00	01	01		
	MT	09	11	00	01	00	00	03	02	03	00
.01	AR	30	23	01	15	00	04	14	16		
	WKPC	09	07	00	02	00	01	07	07		
	MT	42	31	00	07	00	03	17	17	13	00
.03	AR	46	43	04	29	03	19	30	31		
	WKPC	24	20	01	06	01	04	13	15		
	MT	61	50	01	20	01	11	28	32	27	00
.05	AR	55	51	09	36	08	29	37	37		
	WKPC	33	31	02	09	06	06	17	20		
	MT	67	61	03	28	08	22	37	42	38	00
.10	AR	68	64	19	50	23	43	50	52		
	WKPC	45	44	07	20	14	14	30	32		
	MT	77	73	13	45	20	40	54	59	57	00

Table 29. Selected ROC Points for Spuriously High Response Patterns Created from NORC Examinees in the 49-64% Ability Range

False alarm rate	Test	Proportion detected by									
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	SR	DFK
15% Spuriously High Treatment											
.001	AR	00	01	00	00	00	00	00	00		
	WKPC	00	00	00	00	00	00	00	00		
	MT	00	00	00	00	00	00	00	00	01	00
.01	AR	02	03	01	02	00	01	02	02		
	WKPC	01	00	00	00	00	00	01	01		
	MT	02	01	00	01	00	01	03	03	04	00
.03	AR	10	10	02	08	02	03	10	11		
	WKPC	04	02	00	01	03	01	04	02		
	MT	12	08	01	04	01	02	12	09	15	00
.05	AR	18	18	04	15	07	08	15	15		
	WKPC	07	05	01	03	06	03	07	10		
	MT	23	18	01	07	07	07	12	15	21	00
.10	AR	34	34	09	25	14	18	26	25		
	WKPC	20	16	04	10	13	09	14	17		
	MT	43	37	07	16	16	17	24	26	35	00
30% Spuriously High Treatment											
.001	AR	03	01	00	01	00	00	01	00		
	WKPC	00	00	00	00	00	00	00	00		
	MT	01	01	00	00	00	00	01	00	04	00
.01	AR	12	09	00	05	01	01	05	07		
	WKPC	02	01	00	00	00	01	02	02		
	MT	15	10	00	01	00	01	05	06	13	00
.03	AR	26	25	01	16	05	08	18	21		
	WKPC	10	07	01	02	04	02	05	07		
	MT	37	29	01	05	03	06	12	18	28	00
.05	AR	36	36	04	23	10	18	23	25		
	WKPC	17	14	01	03	08	03	08	12		
	MT	48	45	01	09	09	13	18	23	38	00
.10	AR	54	53	13	35	22	29	34	39		
	WKPC	32	30	06	08	16	10	16	21		
	MT	64	62	07	21	21	26	34	38	54	00

Table 30. Selected ROC Points for Spuriously Low Response Patterns Created from NORC Examinees in the 31-48% Ability Range

False alarm rate	Test	Proportion detected by									
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	SR	DFK
15% Spuriously Low Treatment											
.001	AR	00	00	01	01	00	00	00	00		
	WKPC	01	00	00	01	00	00	01	01		
	MT	02	00	00	01	00	00	01	00	00	00
.01	AR	05	01	02	02	00	01	01	01		
	WKPC	09	02	01	07	00	01	06	05		
	MT	11	04	02	05	00	01	05	03	01	00
.03	AR	14	05	08	08	01	05	06	06		
	WKPC	22	15	03	16	00	04	12	14		
	MT	26	17	06	15	00	03	09	10	06	00
.05	AR	19	11	13	12	03	09	09	07		
	WKPC	31	27	09	20	01	05	16	18		
	MT	35	26	14	20	02	08	15	16	10	00
.10	AR	30	23	23	21	13	16	16	16		
	WKPC	46	41	26	35	13	14	26	28		
	MT	52	45	32	33	14	16	27	27	18	04
30% Spuriously Low Treatment											
.001	AR	01	01	01	01	00	00	01	00		
	WKPC	07	01	01	02	00	00	03	03		
	MT	03	03	00	03	00	00	02	01	01	00
.01	AR	11	04	04	05	00	03	03	04		
	WKPC	24	21	02	19	00	02	11	16		
	MT	26	26	07	16	00	02	10	12	06	00
.03	AR	25	16	13	12	01	09	09	10		
	WKPC	41	40	16	31	00	11	20	28		
	MT	50	43	23	30	00	10	17	22	15	01
.05	AR	31	21	20	18	06	14	14	13		
	WKPC	49	48	33	40	03	15	27	32		
	MT	62	54	40	41	05	18	25	30	23	01
.10	AR	48	39	33	31	19	22	22	23		
	WKPC	72	62	55	56	21	30	39	49		
	MT	81	71	63	57	25	33	42	46	33	15

Table 31. Selected ROC Points for Spuriously Low Response Patterns Created from NORC Examinees in the 49-64% Ability Range

False alarm rate	Test	Proportion detected by									
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	SR	DFK
15% Spuriously Low Treatment											
.001	AR	01	00	00	01	00	00	01	00		
	WKPC	07	02	00	01	00	00	01	01		
	MT	08	03	00	01	00	00	02	01	00	00
.01	AR	11	05	01	04	00	01	04	03		
	WKPC	25	19	01	09	00	02	10	08		
	MT	28	24	01	08	00	01	13	10	01	00
.03	AR	23	12	08	12	02	06	11	11		
	WKPC	41	36	04	20	06	09	19	19		
	MT	45	40	08	24	04	06	21	20	05	00
.05	AR	28	20	14	16	09	09	15	13		
	WKPC	47	46	15	26	20	11	25	26		
	MT	54	49	21	32	20	15	29	26	10	00
.10	AR	39	32	25	27	23	10	27	23		
	WKPC	57	58	36	45	38	24	41	40		
	MT	69	64	42	46	42	29	45	43	16	00
30% Spuriously Low Treatment											
.001	AR	02	02	00	01	00	00	01	00		
	WKPC	18	06	00	04	00	00	07	06		
	MT	13	15	01	07	00	01	07	03	01	00
.01	AR	23	10	03	10	00	02	10	08		
	WKPC	40	39	03	28	00	05	23	24		
	MT	50	48	09	28	00	05	27	26	08	00
.03	AR	41	28	20	21	04	11	21	18		
	WKPC	57	55	20	44	02	21	36	41		
	MT	69	62	31	50	02	17	41	44	17	00
.05	AR	47	34	29	21	14	17	25	23		
	WKPC	66	63	37	51	15	25	42	49		
	MT	77	69	53	61	21	33	50	51	24	00
.10	AR	61	49	45	42	33	27	39	36		
	WKPC	78	73	65	67	43	44	57	61		
	MT	87	80	75	73	51	51	66	65	37	02

Table 32. Selected ROC Points for Spuriously Low Response Patterns Created from NORC Examinees in the 65-92% Ability Range

False alarm rate	Test	Proportion detected by									
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	SR	DFK
15% Spuriously Low Treatment											
.001	AR	08	04	00	01	00	00	01	00		
	WKPC	25	11	00	00	00	00	04	01		
	MT	39	24	00	03	00	00	07	02	00	00
.01	AR	33	21	03	11	06	03	12	09		
	WKPC	46	36	00	11	01	03	19	13		
	MT	62	53	05	19	01	06	29	23	02	00
.03	AR	45	34	16	24	23	17	24	23		
	WKPC	61	53	10	24	25	15	31	28		
	MT	74	66	21	41	31	22	42	38	10	00
.05	AR	52	42	26	31	33	24	31	27		
	WKPC	66	61	24	34	43	19	39	37		
	MT	79	72	42	51	55	34	52	46	15	00
.10	AR	61	50	43	45	47	36	44	40		
	WKPC	73	69	49	53	61	38	55	52		
	MT	84	80	68	67	74	50	68	61	25	00
30% Spuriously Low Treatment											
.001	AR	15	14	00	08	00	00	06	01		
	WKPC	40	21	00	07	00	01	20	15		
	MT	44	48	00	29	00	02	32	20	05	00
.01	AR	45	31	07	26	05	10	26	22		
	WKPC	61	61	06	42	00	16	45	42		
	MT	76	73	21	55	00	27	63	57	17	00
.03	AR	58	45	30	41	03	29	44	39		
	WKPC	73	74	30	60	15	41	59	60		
	MT	85	82	51	76	25	50	73	71	32	00
.05	AR	64	51	43	50	36	39	51	45		
	WKPC	80	82	51	69	40	46	67	68		
	MT	90	87	74	81	56	65	80	79	40	00
.10	AR	74	61	61	64	52	52	62	59		
	WKPC	87	84	77	81	67	66	80	80		
	MT	95	92	89	90	80	79	89	88	55	00

Table 33. Selected ROC Points for Spuriously Low Response Patterns Created from NORC Examinees in the 93-100% Ability Range

False alarm rate	Test	Proportion detected by									
		LR_p	LR_z	z_p	z_z	F1	F2	T2	T4	SR	DFK
15% Spuriously Low Treatment											
.001	AR	26	21	00	02	00	00	02	00		
	WKPC	48	27	00	00	00	00	11	04		
	MT	72	56	00	05	00	00	19	04	00	00
.01	AR	56	43	07	14	32	10	19	11		
	WKPC	67	54	02	17	11	05	34	22		
	MT	85	78	15	28	23	14	50	39	01	00
.03	AR	65	56	25	30	55	32	35	31		
	WKPC	76	68	17	32	55	22	47	46		
	MT	91	85	37	57	74	38	66	59	15	00
.05	AR	69	63	38	39	63	42	42	41		
	WKPC	81	72	35	41	70	27	57	54		
	MT	92	88	62	68	87	59	74	68	29	00
.10	AR	73	67	57	60	74	57	59	57		
	WKPC	85	78	61	65	81	46	75	72		
	MT	94	90	84	84	94	73	88	84	45	00
30% Spuriously Low Treatment											
.001	AR	39	37	00	19	00	00	13	04		
	WKPC	62	43	00	10	00	05	41	30		
	MT	78	77	01	50	00	10	65	46	06	00
.01	AR	72	55	09	45	23	26	44	43		
	WKPC	78	75	13	53	01	29	68	64		
	MT	96	91	36	78	04	58	88	83	37	00
.03	AR	81	66	45	61	58	55	64	61		
	WKPC	85	83	40	72	34	59	81	80		
	MT	98	94	74	92	64	80	93	92	64	00
.05	AR	83	70	60	70	69	64	70	67		
	WKPC	89	85	63	81	62	66	85	85		
	MT	98	95	88	96	87	90	95	94	73	00
.10	AR	88	77	78	83	81	77	80	79		
	WKPC	94	89	85	89	83	83	92	92		
	MT	99	97	98	98	95	95	98	98	84	00

In sum, the distributions of index scores for the NORC normals had more extreme values than did the distribution for the Study One simulated normals. Detection rates of spuriously high examinees did not significantly decrease, however, because there were comparable shifts in the distributions of index scores for the aberrant samples.

The results for the spuriously low conditions are shown in Tables 30 through 33. The detection rates for LR_p are somewhat lower in these tables than the comparable rates (shown in Tables 22 through 25) obtained with simulated data. The rates for LR_1 and z_1 remained basically unchanged. The detection rates for LR_p decreased for two reasons. First, as noted above, the distributions of index scores for the NORC normals shifted toward more extreme values. Second, the distributions of LR_p scores for the spuriously low conditions were essentially unchanged. Thus, the "signal" was unchanged but the "noise" increased; therefore, the signal-to-noise ratio decreased.

Although the rates of detection of spuriously low response patterns were lower for LR_p with the NORC data than with the simulated data, some impressive detection rates were nonetheless obtained. For example, LR_p detected 85% and 62% of the 15% spuriously low examinees for the very high and high ability ranges at a 1% false alarm rate. The corresponding rates were 96% and 76% for the 30% spuriously low treatment.

Discussion

The transition from simulated data in Study One to real data in Study Two was very successful for the three-parameter logistic appropriateness indices. Although detection rates for LR_p tended to be lower with the real data, some impressive results were nonetheless obtained. For example, 82% of the NORC examinees in the lowest ability range who were subjected to the 30% spuriously high treatment could be detected by the optimal LR_p index when the false alarm rate was 1%; 75% could be detected by z_1; and 64% could be detected by T2.

In contrast to the high detection rates obtained by the IRT appropriateness indices, very low detection rates were obtained by the SR measure. For example, only 4% of the very low ability, 30% spuriously high response patterns were identified by SR at a 1% false alarm rate. The results for SR are, in fact, even worse than they appear: SR is based on 30 AR items, 50 WKPC items, 25 MK items, and 25 GS items. Thus, a total of 130 items were used for SR. The IRT appropriateness indices used only 80 items; considerably higher detection rates would be expected if all 130 items were used.

The transition from simulated to real data was less successful for the LR_p index in the spuriously low conditions. Detection rates were lower for the real data because the distributions of index scores for normal NORC examinees were shifted toward more extreme values. The distributions for spuriously low response vectors, in contrast to the spuriously high response patterns, were not similarly shifted.

One hypothesis about the differences between the results for the real and simulated data sets concerns the distributions of ability. For simulated data, abilities were distributed as bivariate normal, with zero means, unit variances, and a correlation of .8. The distributions of ability for the real data were clearly nonnormal: A second mode of the density was evident at $\theta = -5$. This second mode is clearly shown in Figure 4.

Why would a second mode appear at a very low ability? Since the NORC examinees were not a sample of actual recruits, it is possible that some were poorly motivated to do their best. Indeed, some examinees omitted every item on entire tests. Thus, we are led to hypothesize that the bivariate ability distribution contains a nontrivial point mass corresponding to examinees who were very poorly motivated. An optimal index for spuriously low examinees based on the estimated distribution of ability should lead to increased rates of detection.

V. DISCUSSION

In the present effort, several new appropriateness indices were developed. These indices, as well as a number of appropriateness indices previously developed, were carefully evaluated in a series of studies. By comparing detection rates to the rates obtained by the optimal apropriateness indices developed by Levine and Drasgow (1984; 1987), we were able to determine the effectiveness of all of the indices in an absolute sense. Detection rates for the three best practical indices (z_s, T2, and T4) are presented in Figure 8 as percentages of the optimal detection rate (at a 1% error rate).

A major result of this effort is the finding that a few of the practical appropriateness indices (namely, z_s, T2, and T4) effectively detect aberrant response patterns across a fairly wide range of conditions. Multi-test extensions of these indices were developed for situations in which examinees complete a battery of short unidimensional tests. The multi-test extensions of z_s, T2, and T4 were found to provide high rates of detection of aberrant response patterns when simulated and actual ASVAB data were used. Thus, it was concluded that these indices, which are all based on IRT, are strong candidates for use in operational settings.

The standardized residual (SR) index provides another approach to the detection of inappropriate response patterns. Unlike IRT indices, which analyze the internal consistency of a response pattern, the SR index uses external information such as scores on other tests. This external evidence is used to predict scores on the tests of interest (e.g., AFQT subtests), and large errors of prediction are taken as indicating that test scores are aberrant.

The SR index, in contrast to the IRT indices, was found to [illegible] under all conditions. It therefore seems to be a weak operationalization of an important idea. IRT provides a much more precise and powerful method of detecting aberrant response patterns than the classical test theory concepts used by SR.

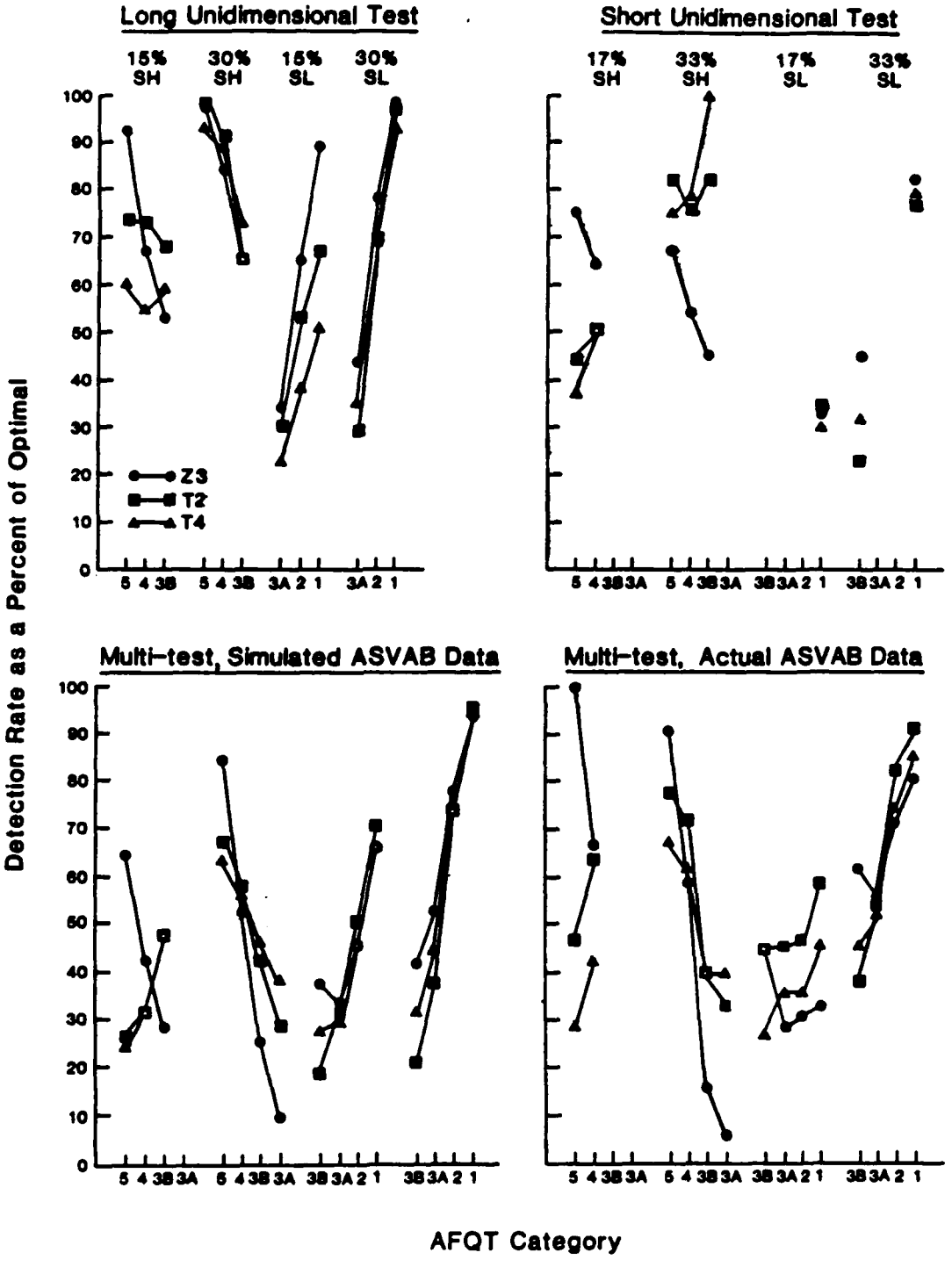


Figure 8. Detection rates of z_3, T2, and T4 expressed as proportions of the rate of the optimal index at a 1% false alarm rate. (Rates are not plotted when the optimal index detected less than 10% of the aberrant sample.)

How effective are the best practical appropriateness indices in relation to optimal indices? The practical appropriateness indices are much better than non-IRT alternatives such as the SR measure, but sometimes fall short of optimal. Therefore, it seems that operational use of z_3, T2, and T4 is justified. Moreover, a program of research designed to develop and validate better practical appropriateness indices is also warranted. This conclusion was reached because z_3, T2, and T4 decisively outperformed SR and other IRT indices, but fell short of optimality in some cases.

The optimal appropriateness indices used in the present research seem to be simultaneously too specific and not specific enough to use as practical appropriateness indices. They are too specific in that different optimal indices must be computed for differing percentages of spuriously high and spuriously low responses. They are not specific enough in that ability is assumed to be distributed as standard normal in both the normal and aberrant groups. More specific assumptions about ability distributions, particularly for the aberrant group, would seem to be desirable in many situations.

Therefore, it is important to develop a "second generation" of optimal indices that could be used in practice to test hypotheses that are very general in some ways but very specific in others. Examples of some hypotheses that may be important to test include the following:

1. Was a response vector generated by a normal examinee or was it generated by a very low ability (AFQT Category V) examinee who was cheating on 10 to 30 items? Low ability cheaters would be expected to have high rates of attrition in training and generally poor on-the-job performance, both of which are very costly.

2. Was a response vector generated by a high average (AFQT Category 3A) examinee or a low average (AFQT Category 3B) examinee who was cheating on a moderate number of items? Recruitment bonuses for AFQT Category 3A scores may provide a powerful incentive for examinees slightly below average to cheat.

3. Suppose it is known that part or all of one subtest is no longer secure. Was a response pattern generated by a normal examinee or by an examinee who had prior access to the compromised items?

4. Are members of an ethnic minority penalized because a test was developed and standardized using majority group members as examinees? The likelihood of the response pattern could be computed using item parameters estimated from a majority group sample and from a minority group sample. If the test is fair, then even the optimal appropriateness index would be unable to effectively classify majority and minority group members. In this way, the methodology of optimal indices is applied to determine the extent to which ethnicity can be determined from item response patterns.

Refinements in optimal indices would enable very powerful detection of aberrant response patterns. For example, suppose we suspect that a very low ability examinee has been given answers to a moderate number of items on the AR, WK, and PC subtests in order to obtain an AFQT score that qualifies him/her for a bonus. Furthermore, suppose that there was no cheating on the non-AFQT subtests. Then we could test the hypothesis that the examinee was

normal against the hypothesis that a low ability examinee cheated on 20 to 30 items on the AR, WK, and PC subtests and cheated on 0 items on the MK and GS tests. Examinees who are aberrant in this particular way should be clearly identifiable.

A significant part of the theory necessary for more sophisticated optimal indices has already been developed by Levine and Drasgow (1984; 1987). Nonetheless, a considerable amount of work is necessary to transform their theoretical notions, which were developed in the context of a unidimensional latent trait space, into methods that can be used to test the aberrance hypotheses listed above.

It may seem that computing second-generation optimal indices would be extremely burdensome. It is true that extensive calculations would be necessary. The recursive methods described by Levine and Drasgow (1984; 1987) and the quadratic approximation and multi-test generalizations developed here considerably reduce the computing load. Furthermore, the rapid advances in Levine's (1985a; 1985b) MFS theory allow algebraic simplifications and eliminate the need for arbitrary assumptions about the ability density. In particular, MFS now permits one to bypass the quadratic approximation used in Chapter IV and relax the assumption of multivariate normal abilities. Multidimensional extensions of Levine's theory are being developed to estimate the joint distribution of several abilities.

Finally, there are two important substantive questions about Appropriateness Measurement that need to be addressed. First, the ability densities estimated from the NORC sample depart significantly from a normal density. This has led us to reconsider the way in which we compute optimal indices. However, the NORC sample is not a sample of individuals who are actually trying to enlist in the military. Would our results concerning ability densities be replicated if data from actual recruits were used? Or would the results be more similar to our studies with SAT-V data?

The second substantive question concerns the distributions of appropriateness index scores in samples of women and ethnic minorities. Finding similar distributions across all relevant groups would support the view that standardized tests in general, and the ASVAB in particular, assess ability fairly. This finding would be highly significant in light of the underprediction of women's performances reported in some military training schools (Dunbar & Novick, 1985).

REFERENCES

Bock, R. D., & Mislevy, R. J. (1981). The Profile of American Youth: Data quality analysis of the Armed Services Vocational Aptitude Battery. Chicago: National Opinion Research Center.

Chung, K. L. (1974). A course in probability theory (2nd ed.). New York: Academic Press.

Cook, R.D., & Weisberg, S. (1982). Regression and influence in regression. New York: Chapman and Hall.

Cliff, N. (1979). Test theory without true scores? Psychometrika, 44, 373-393.

Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. Journal of Applied Psychology, 72, 10-18.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. Applied Psychological Measurement, 10, 59-67.

Drasgow, F., Levine, M.V., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.

Dunbar, S.B., & Novick, M.R. (1985). On predicting success in training for males and females: Marine Corps clerical specialties and ASVAB Forms 6 and 7 (Technical Report 85-2). Iowa City, IA: CADA Research Group, 356 Lindquist Center, University of Iowa.

Efron, B., & Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika, 65, 457-487.

Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. Hambleton (Ed.), Applications of item response theory. Vancouver: ERIBC.

Hulin, C.L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Levine, M.V. (1983). The trait in latent trait theory. In D.J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Levine, M.V. (1985a). Classifying and representing ability distributions (Measurement Series 85-1). Champaign, IL: Model-Based Measurement Laboratory, Department of Educational Psychology, University of Illinois.

Levine, M.V. (1985b). Constrained maximum likelihood estimation of ability distributions (in preparation). Champaign, IL: Model-Based Measurement Laboratory, Department of Educational Psychology, University of Illinois.

Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42-56.

Levine, M.V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.

Levine, M.V., & Drasgow, F. (1984). Performance envelopes and optimal appropriateness measurement (Measurement Series 84-5). Champaign IL: Model-Based Measurement Laboratory, Department of Educational Psychology, University of Illinois.

Levine, M. V., & Drasgow, F. (1987). Optimal appropriateness measurement. Psychometrika, 52, in press.

Levine, M.V., & Rubin, D.F. (1979). Measuring the appropriateness of multiple choice test scores. Journal of Educational Statistics, 4, 269-290.

Levine, M. V., & Williams, B. (1985). Methods for estimating ability densities (in preparation). Champaign, IL: Model-Based Measurement Laboratory, Department of Educational Psychology, University of Illinois.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addision-Wesley.

Massey, F. J., Jr. (1951). The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 46, 68-78.

Mislevy, R. J., & Bock, R.D. (1983). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D.J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), The handbook of social psychology (2nd ed.). Reading, MA: Addison-Wesley.

Ree, M. J., Mullins, C. J., Mathews, J. J., & Massey, R. H. (1982). Armed Services Vocational Aptitude Battery: Item and factor analyses of Forms 8, 9, and 10 (AFHRL-TR-81-55, AD-113 465). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Rudner, L. M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

Sato, T. (1975). The construction and interpretation of S-P tables. Tokyo: Meiji Tosho. (in Japanese).

Swanson L. & Foley, P. Development of Armed Forces Qualification Test (AFQT) Deliberate Failure Keys for Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10 (NPRDC Special Report 83-4) San Diego, CA: Navy Personnel Research and Development Center.

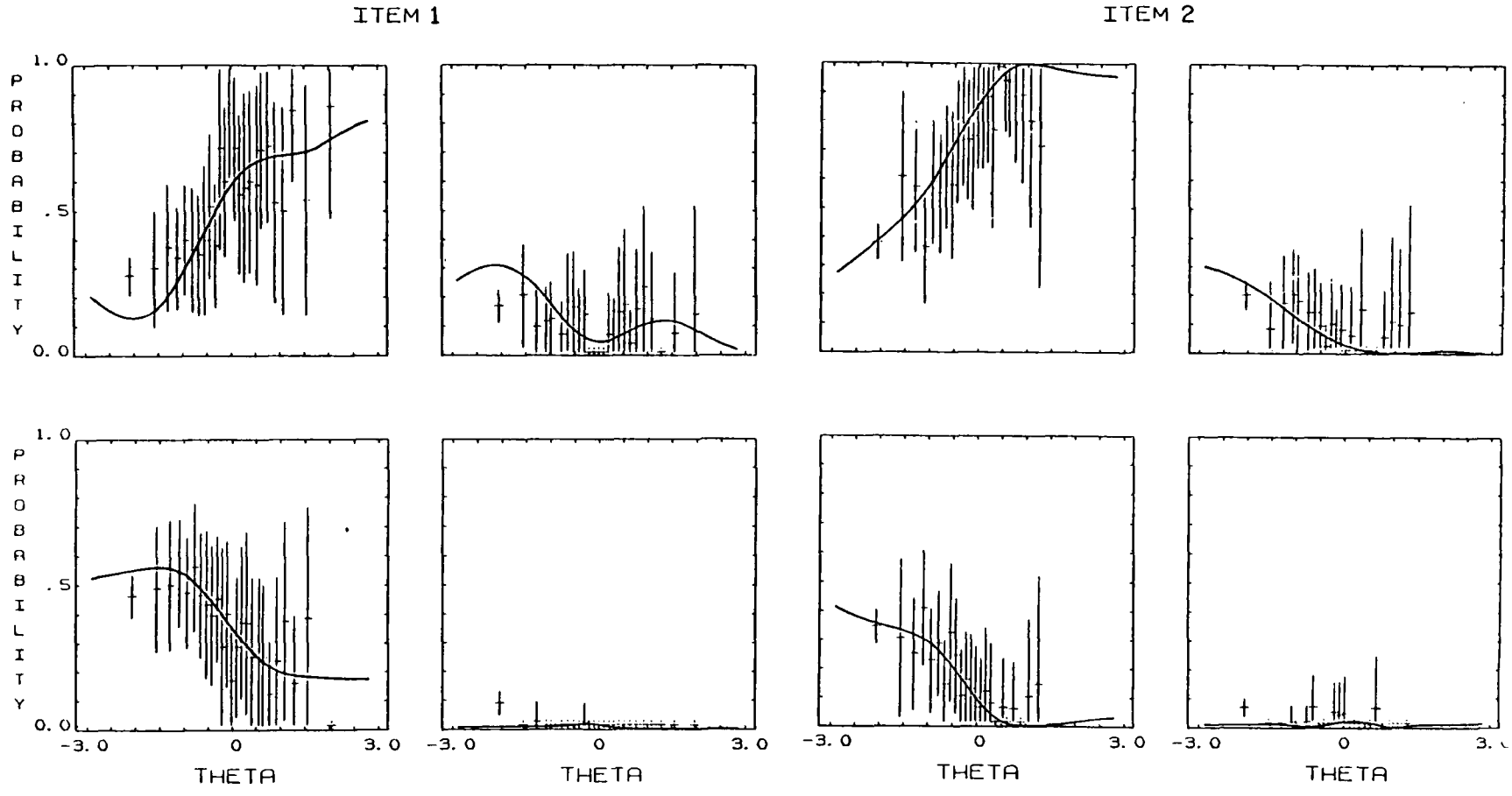
Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

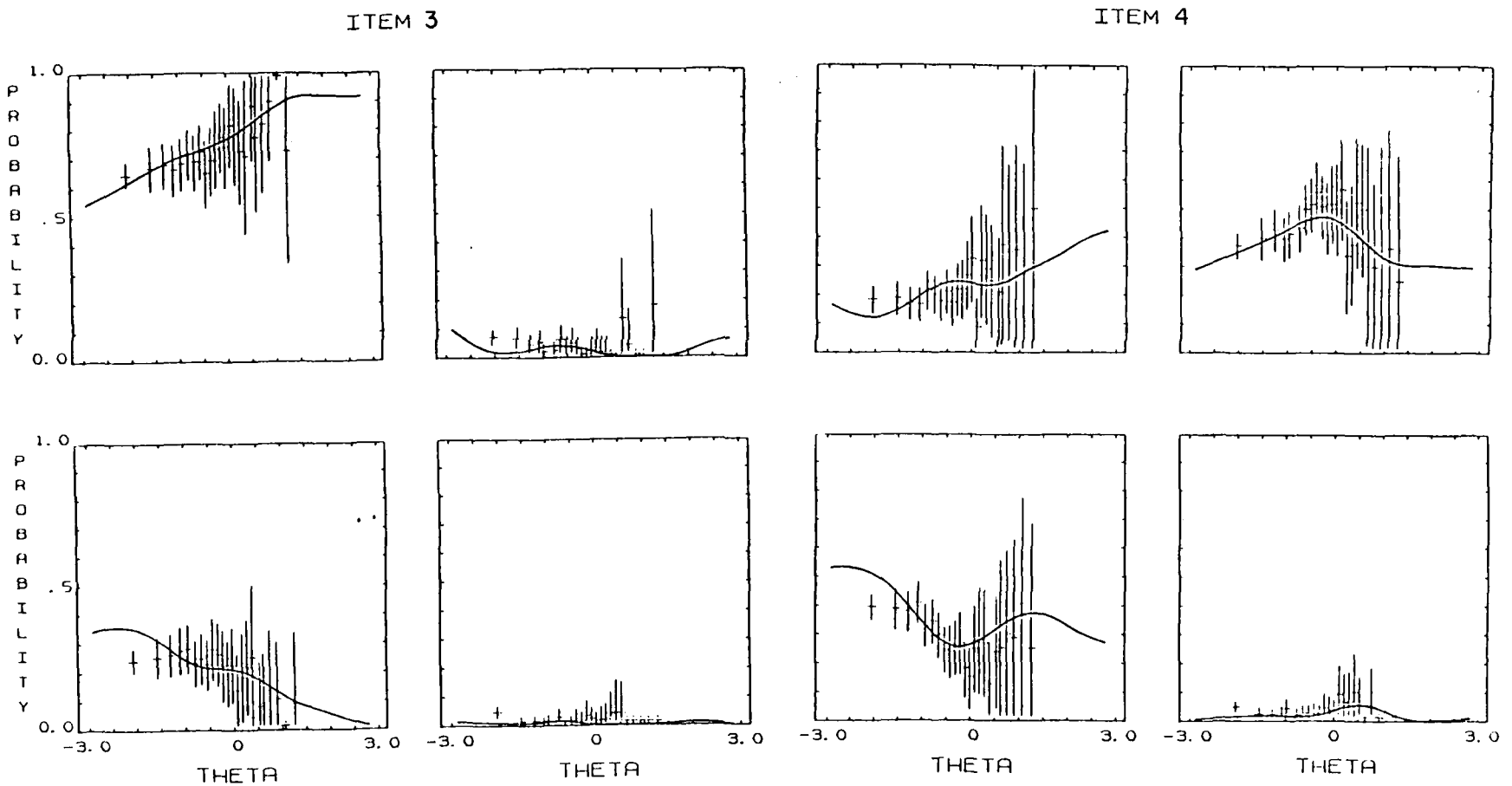
Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual response patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7, 81-96.

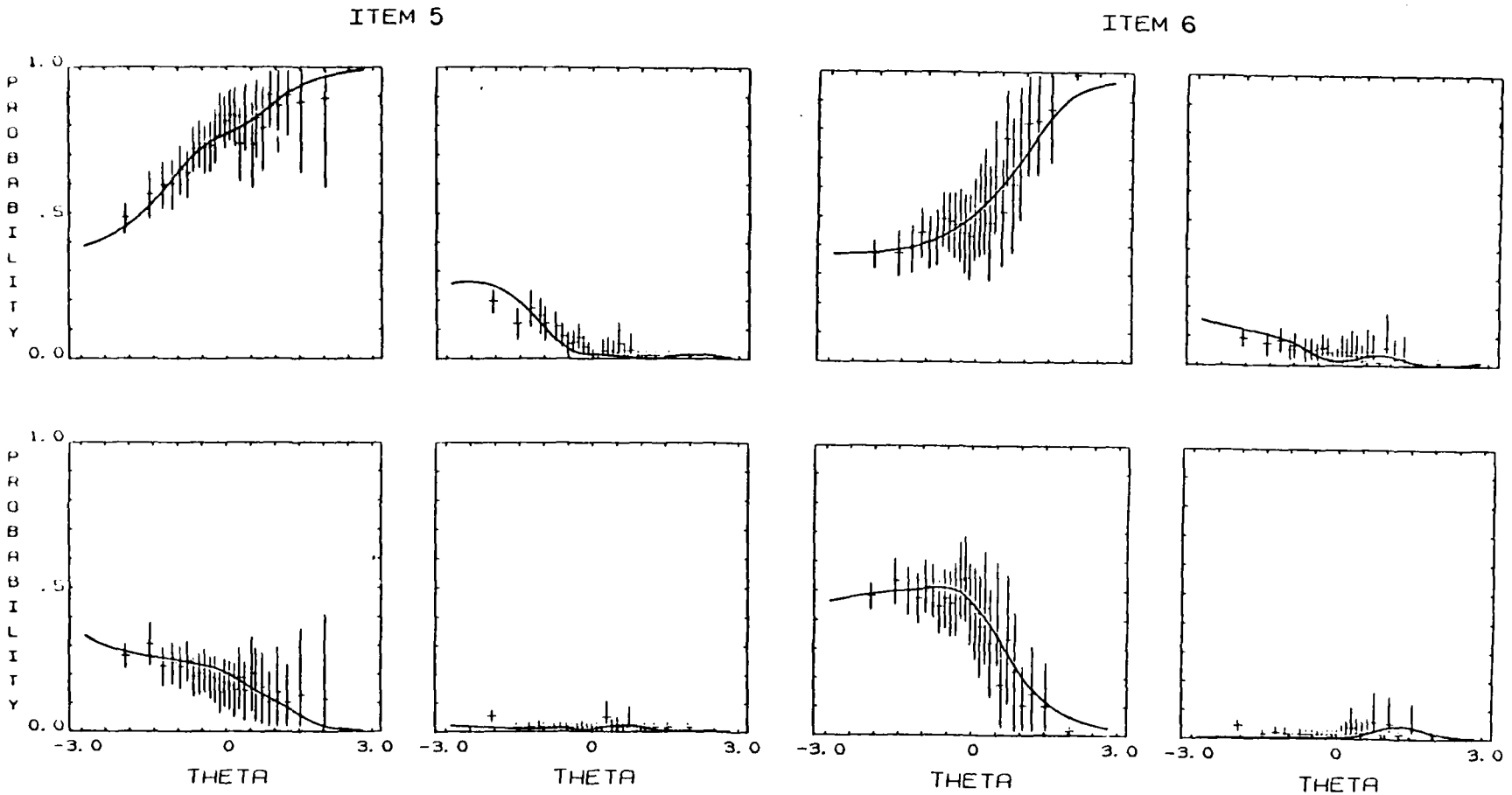
Wood, R.L., Wingersky, M.S., & Lord, F.M. (1976). LOGIST - A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service.

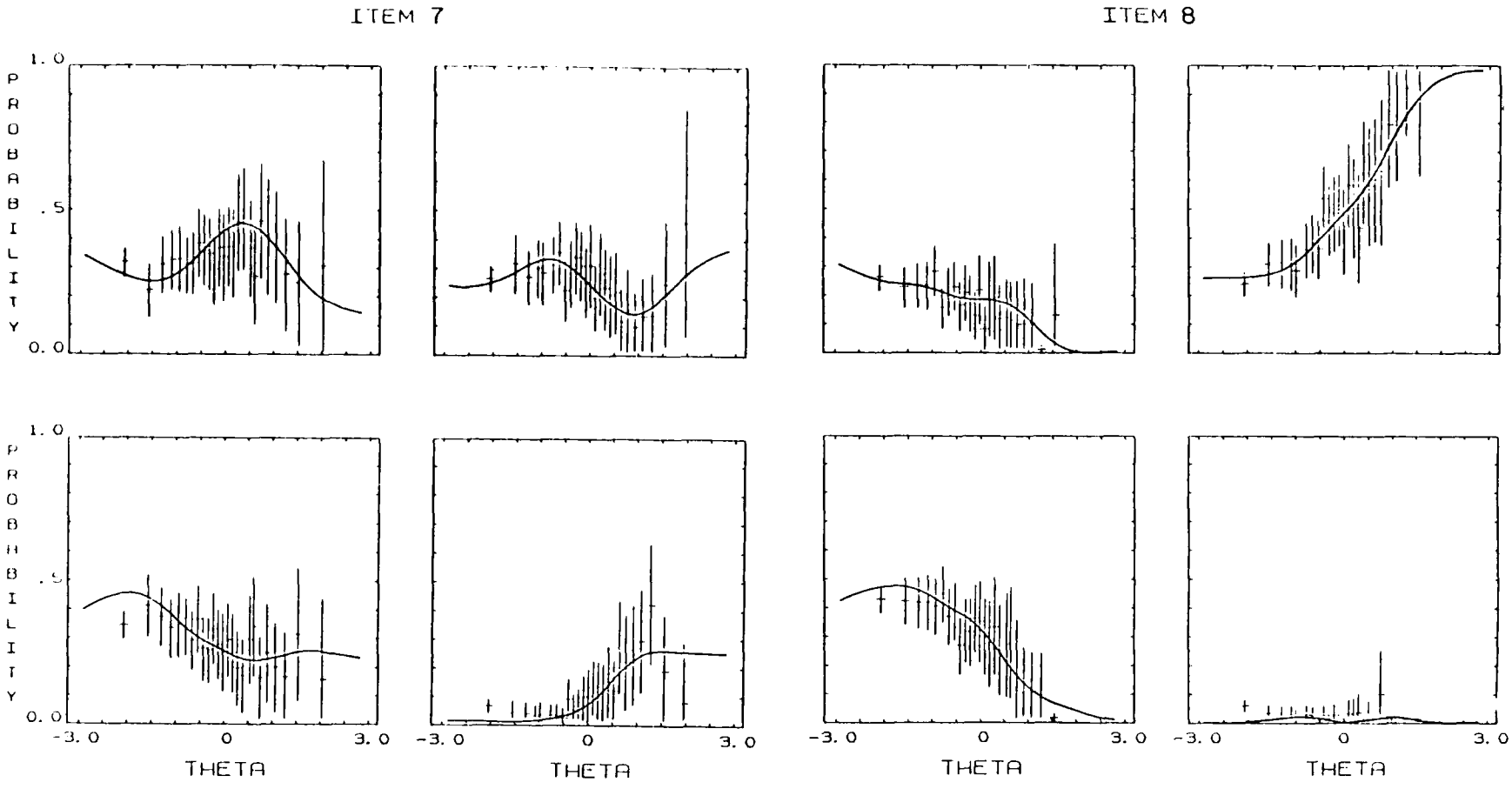
Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

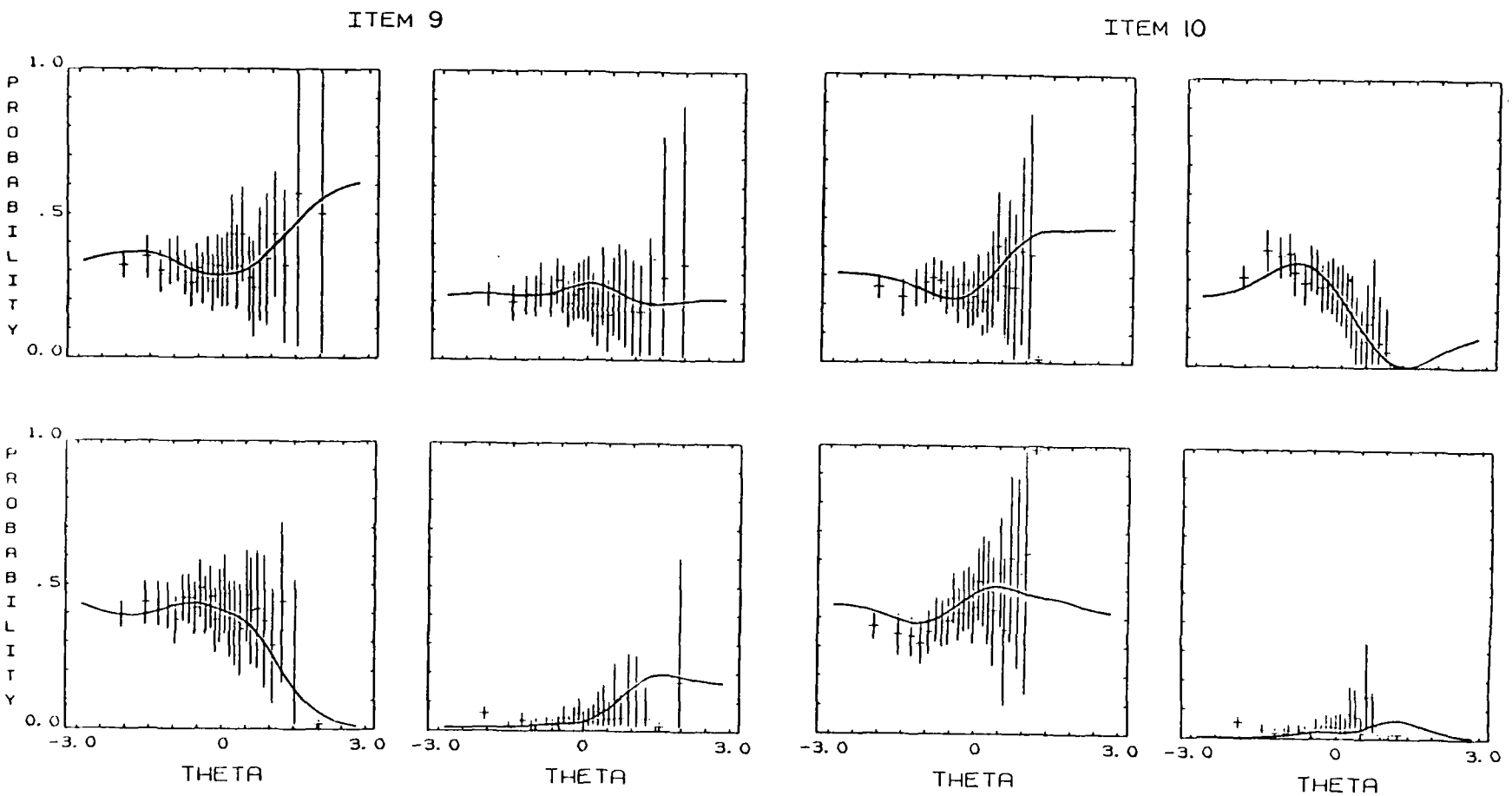
APPENDIX A: GOODNESS OF FIT OF AR COCCs ESTIMATED FROM A SAMPLE OF N=2,891 AND EVALUATED USING THE ENTIRE SAMPLE OF N=11,914

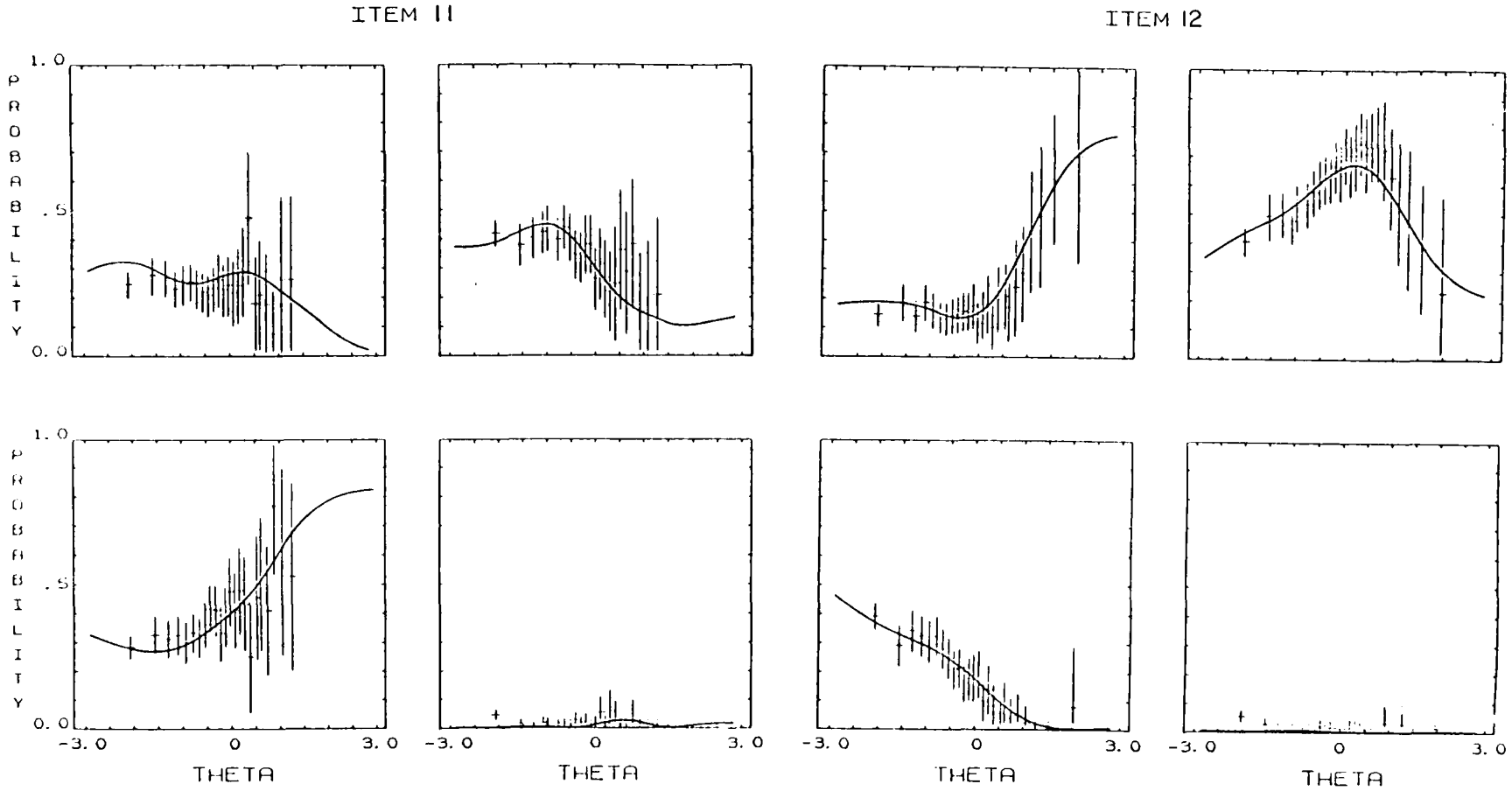

ITEM 1
ITEM 2
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

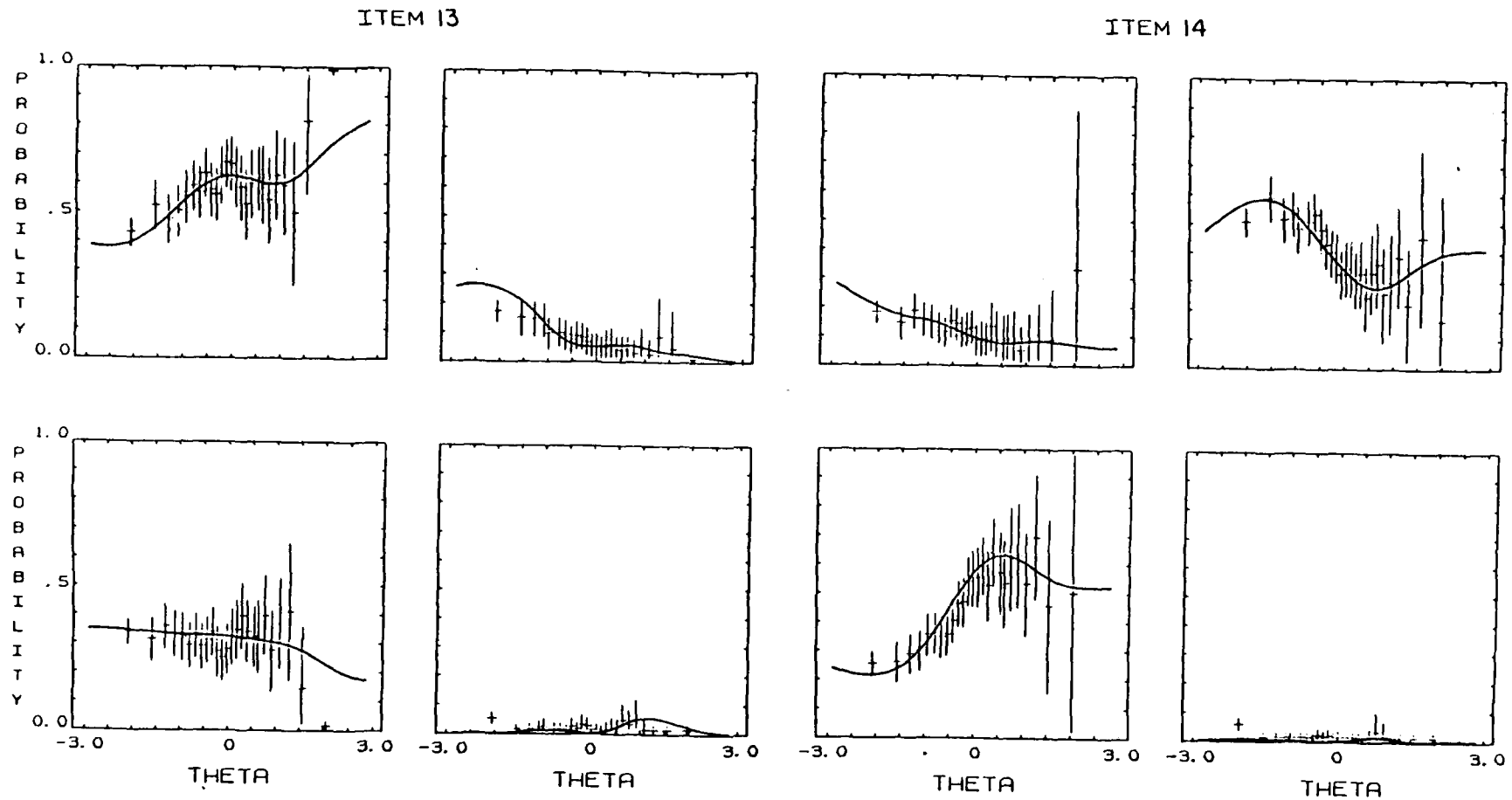

ITEM 3
ITEM 4
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

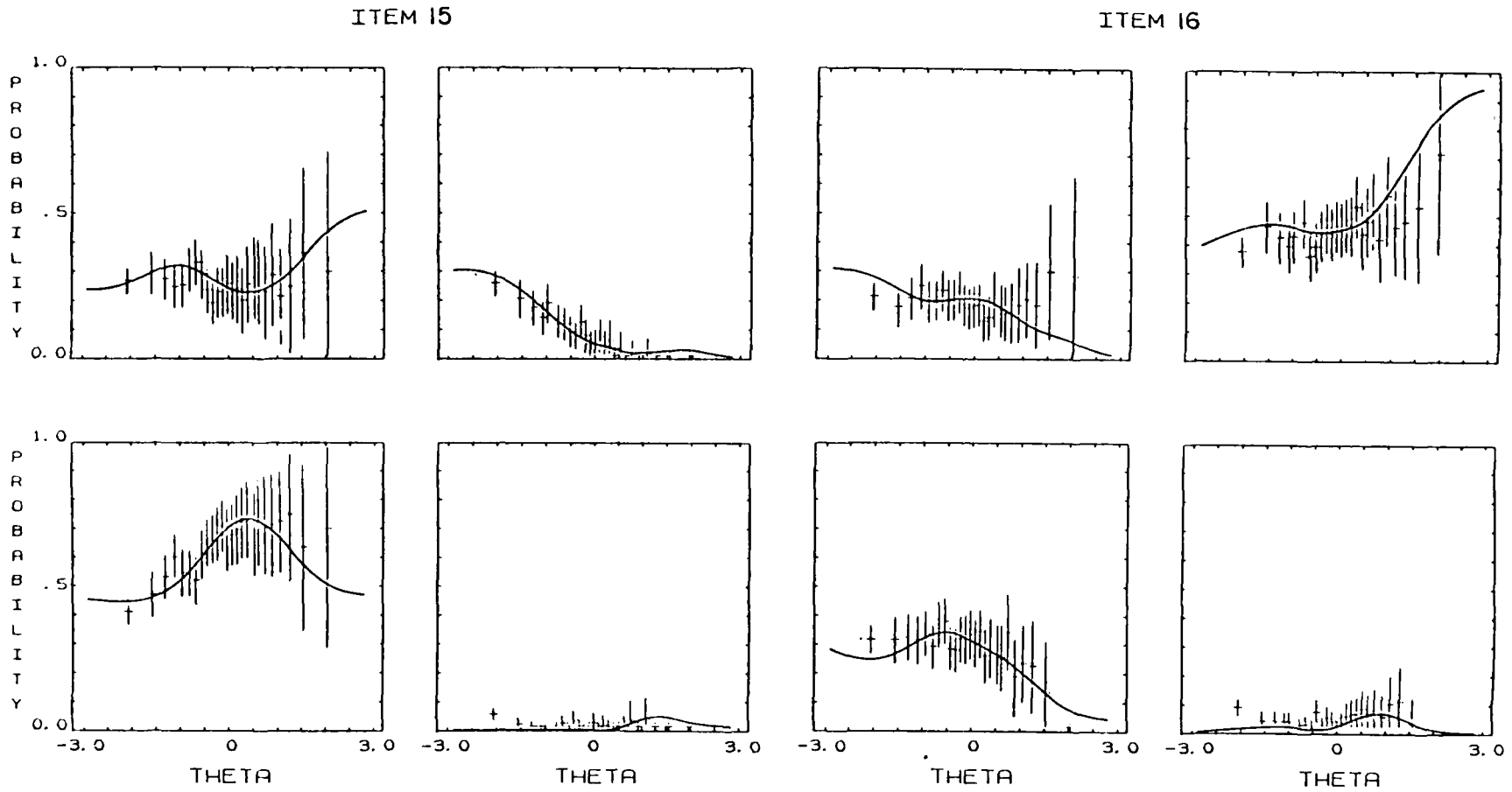

ITEM 5
ITEM 6
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

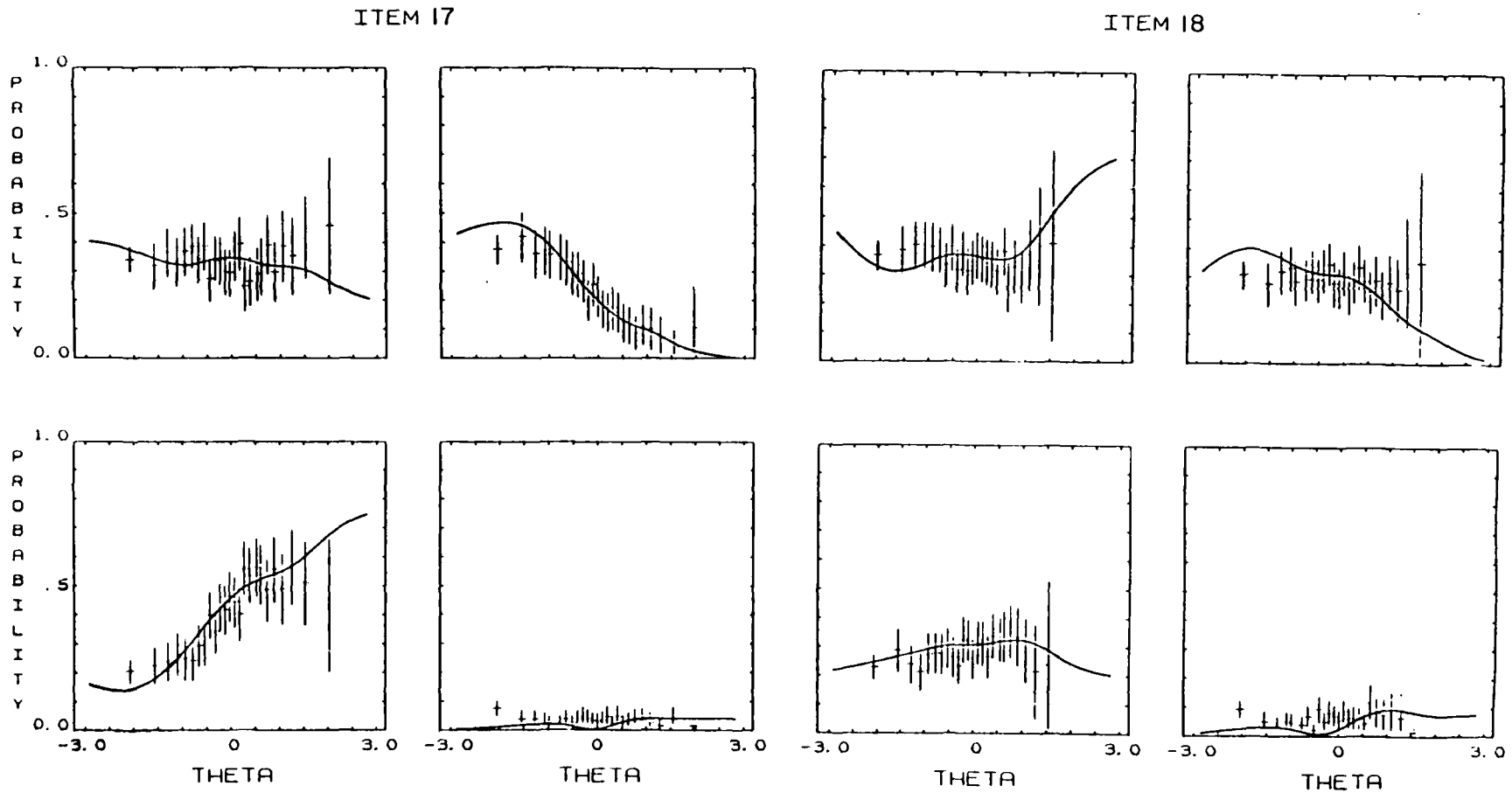

ITEM 7
ITEM 8
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

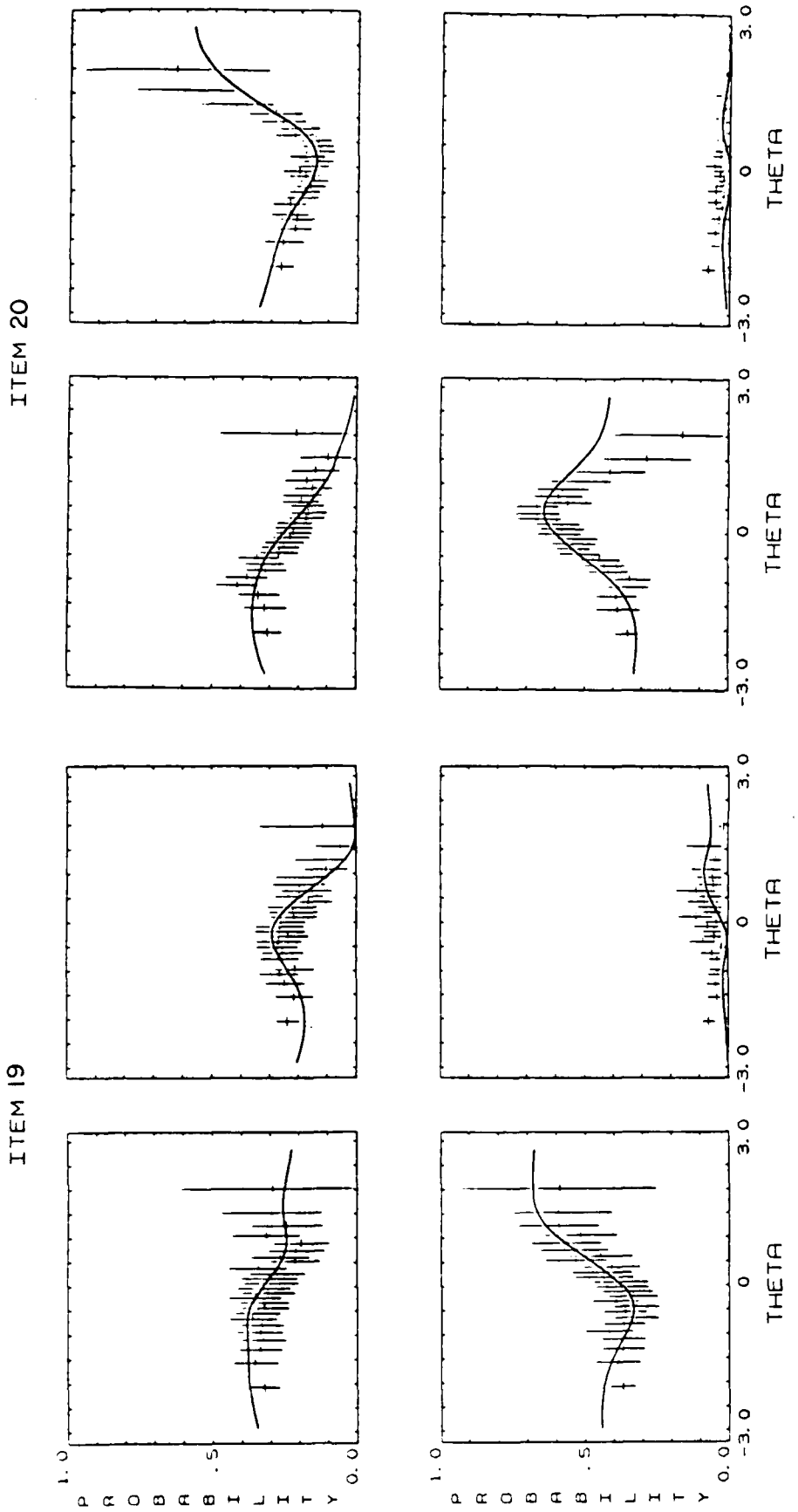

ITEM 9
ITEM 10
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

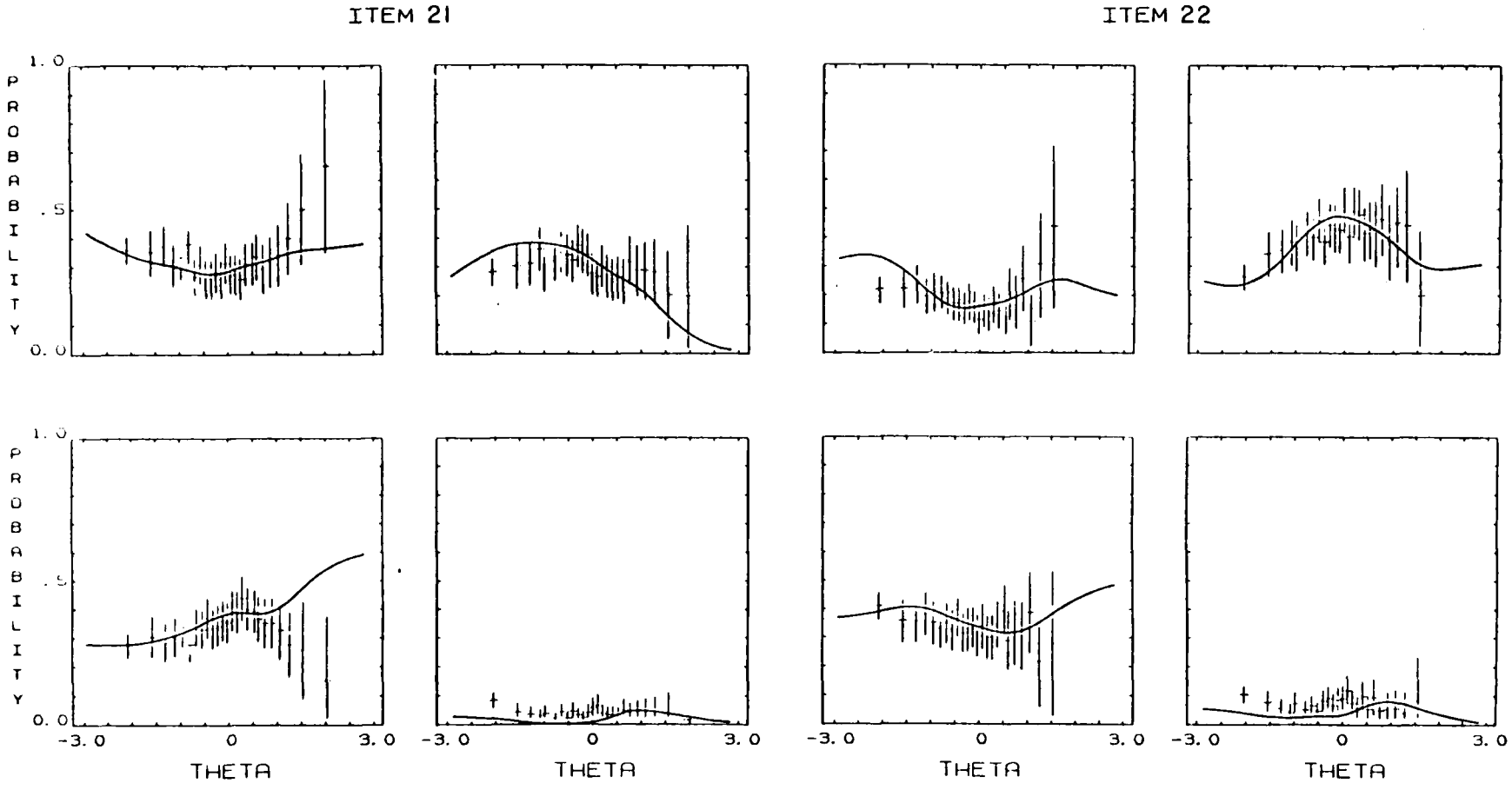

ITEM 11
ITEM 12
1.0
.5
0.0
PROBABILITY
-3.0
0
3.0
THETA

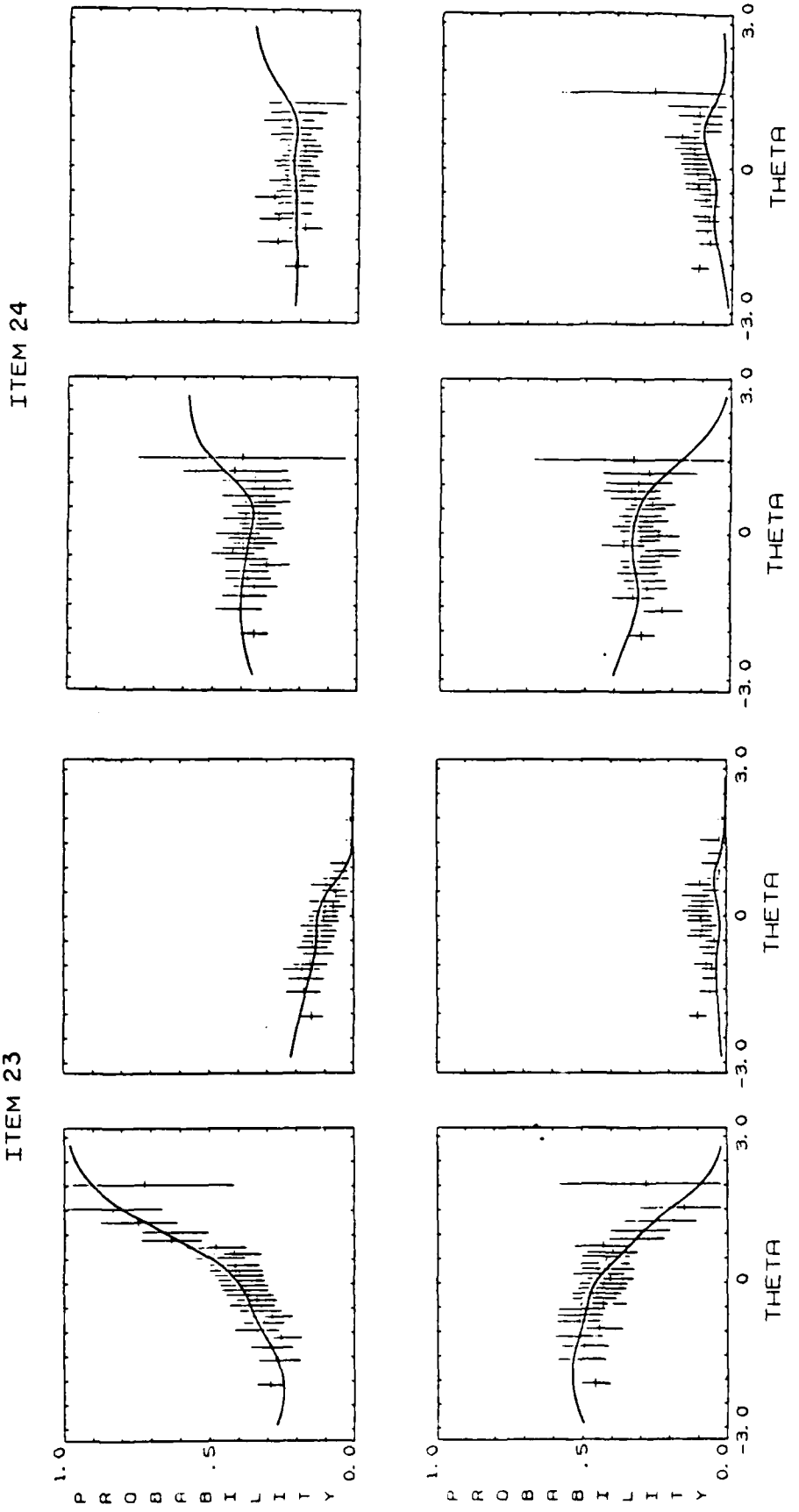

ITEM 13
ITEM 14
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

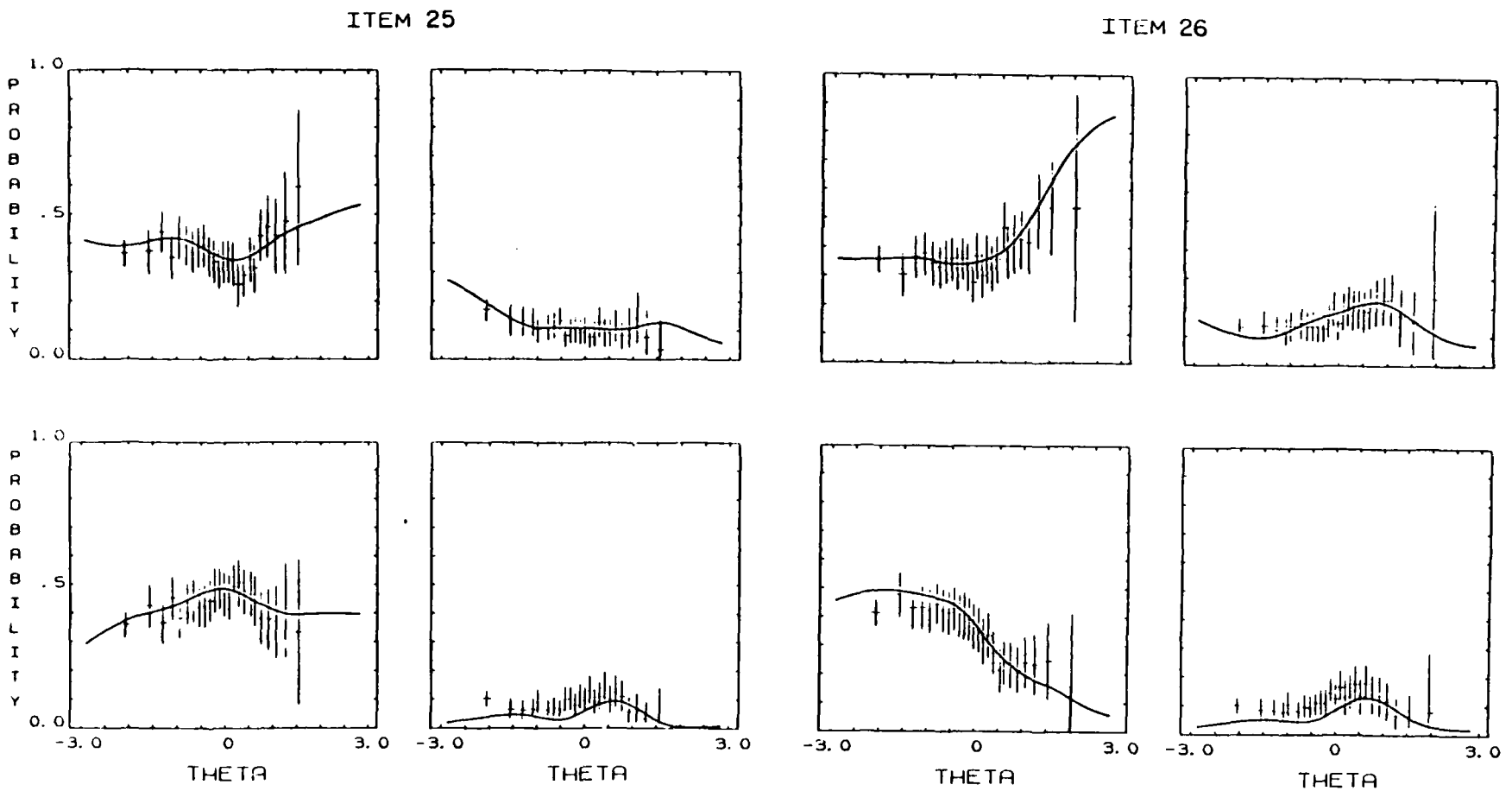

ITEM 15
ITEM 16
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

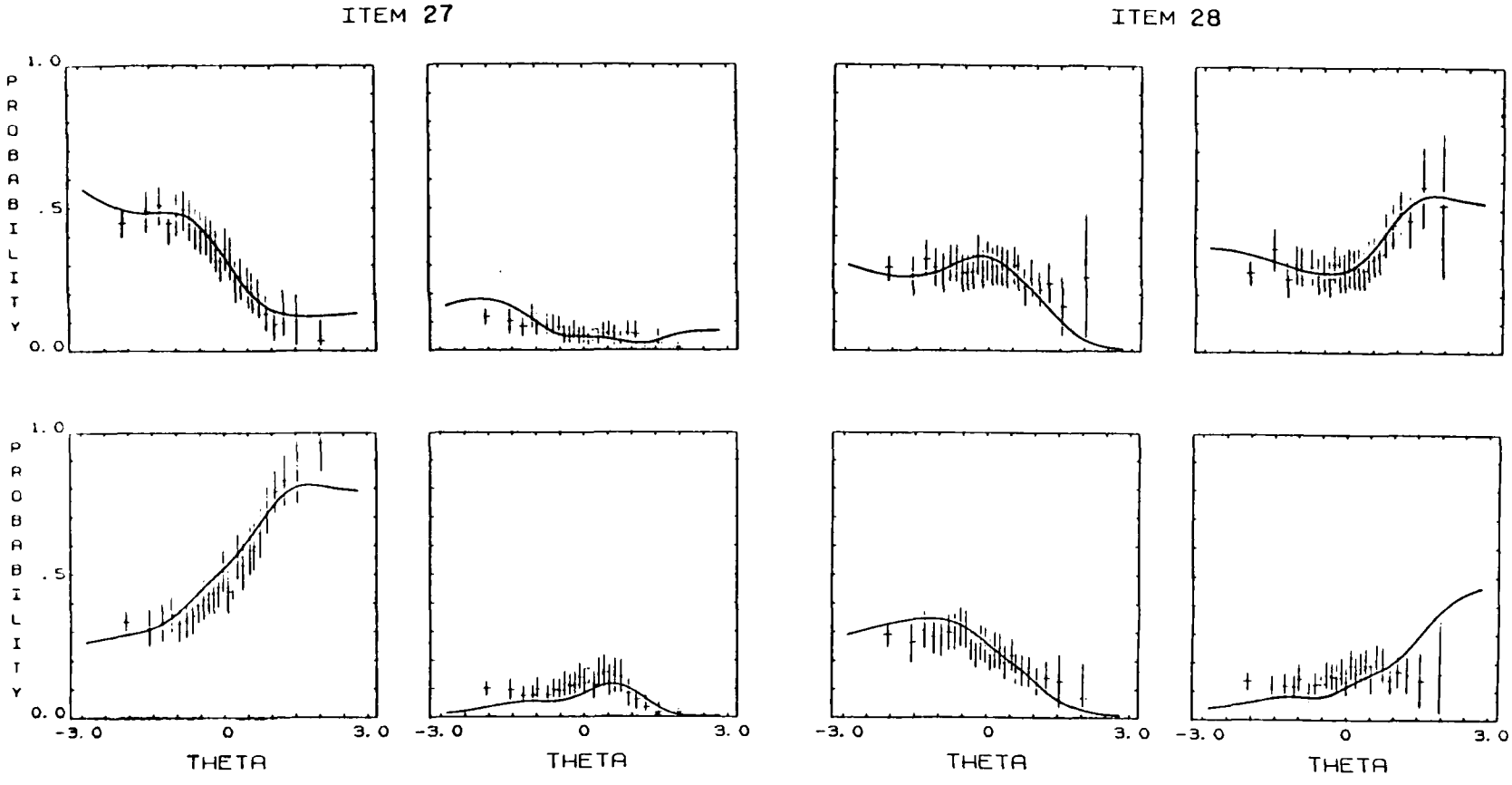

ITEM 17
ITEM 18
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

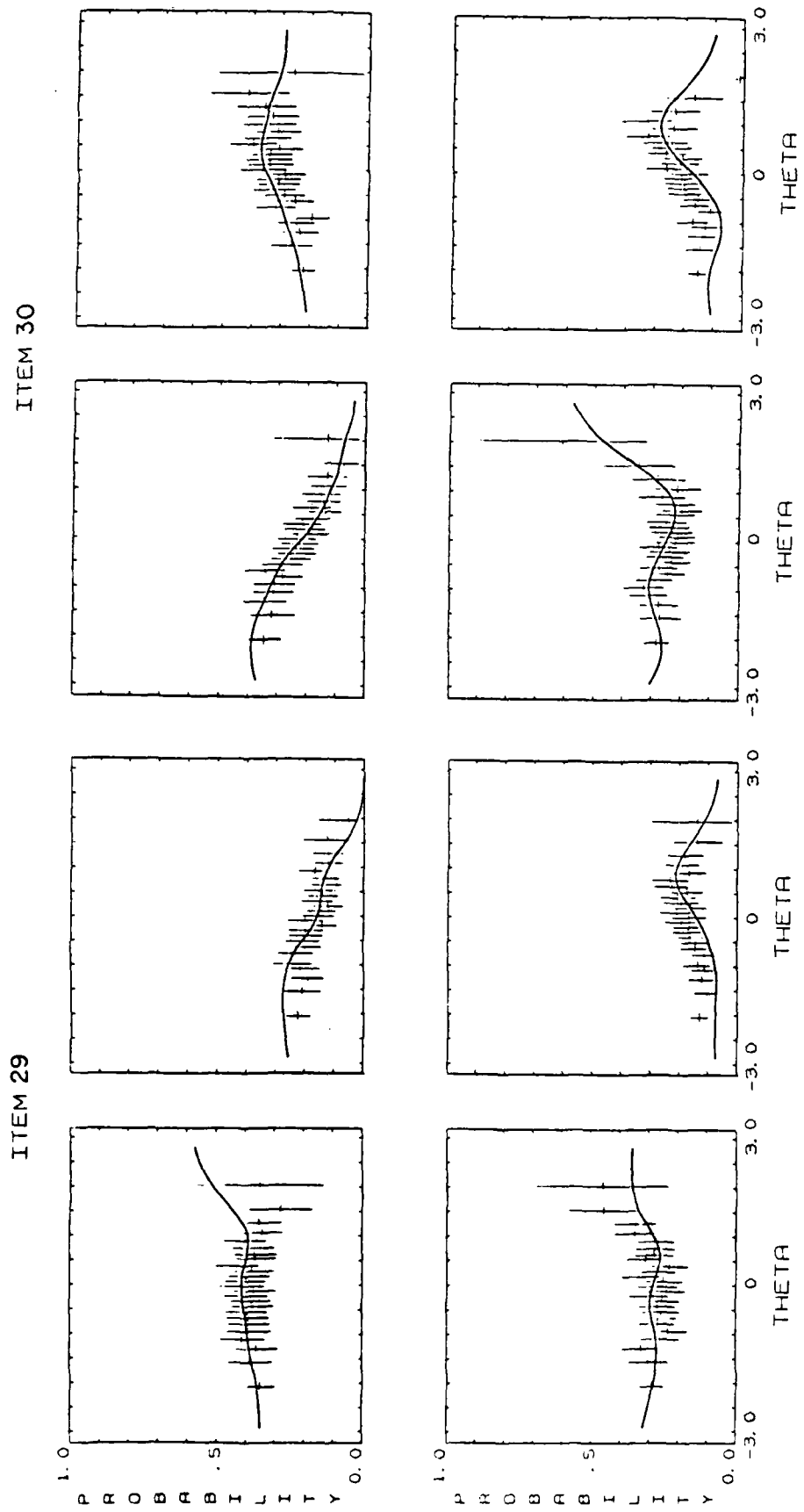

ITEM 19
ITEM 20
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA


ITEM 21
ITEM 22
1.0
.5
0.0
PROBABILITY
-3.0
0
3.0
THETA

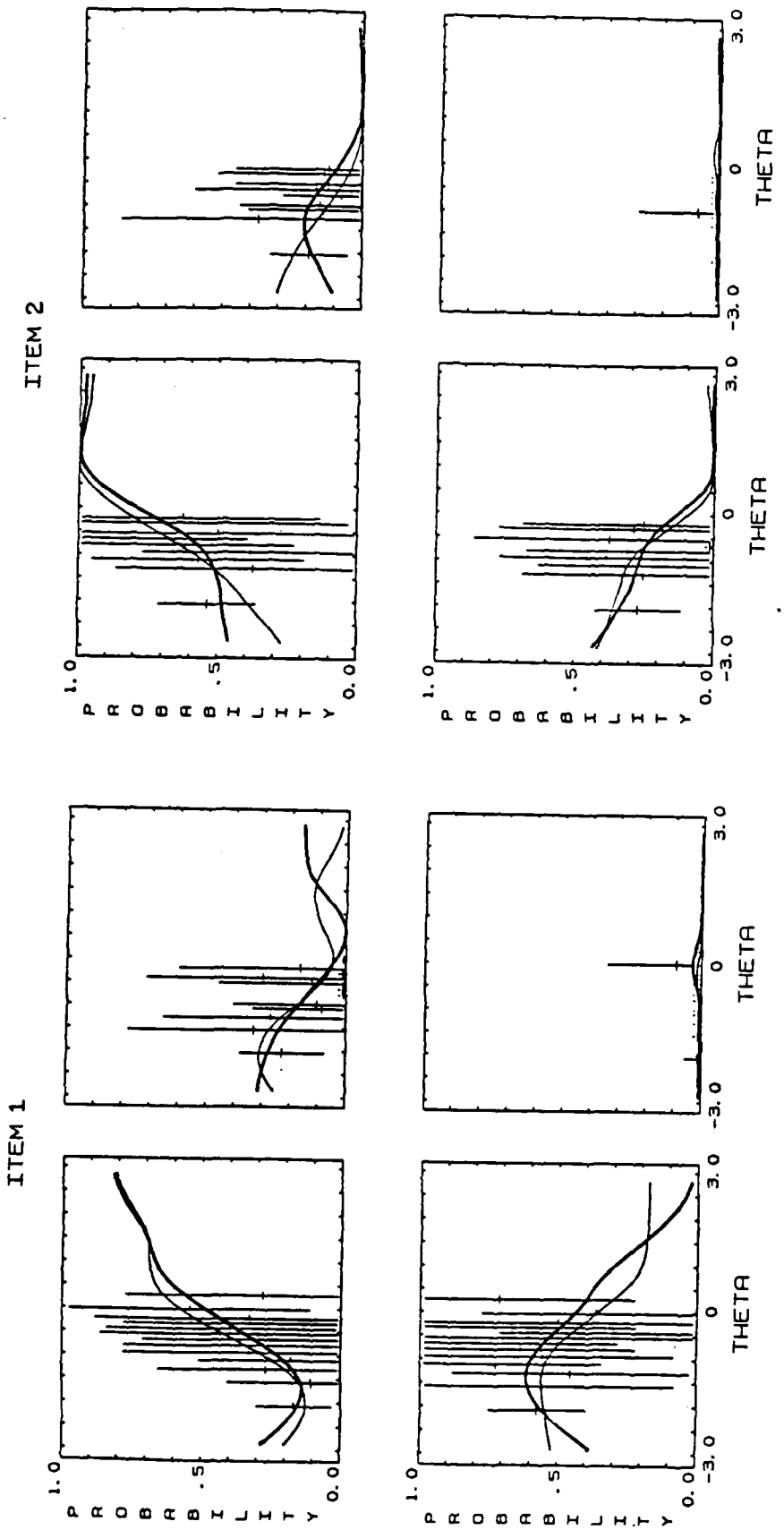

ITEM 23
ITEM 24
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA


ITEM 25
ITEM 26
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

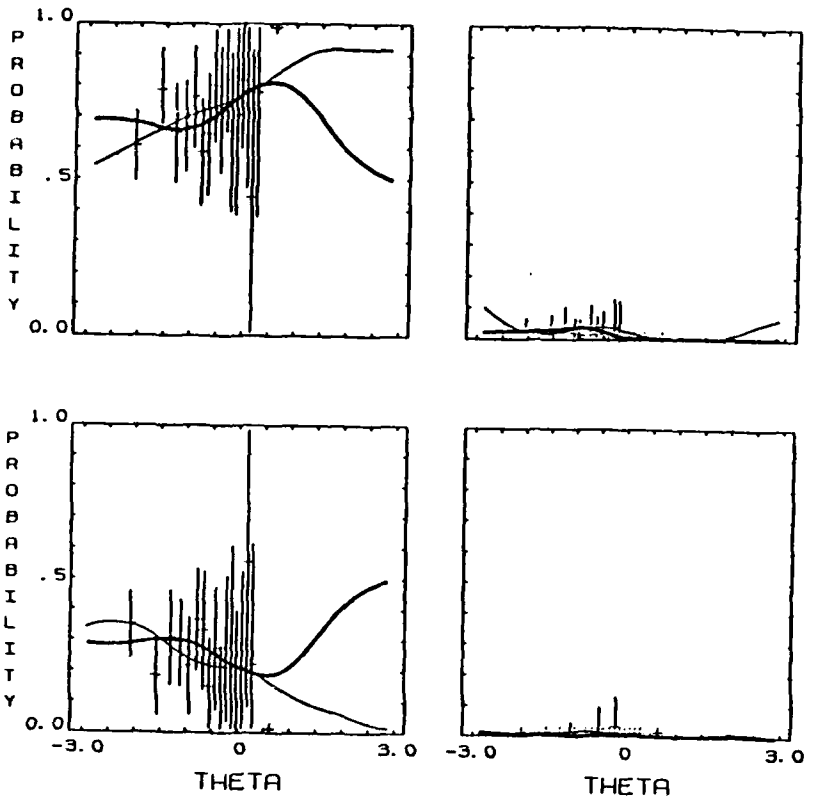

ITEM 27
ITEM 28
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA


ITEM 29
ITEM 30
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

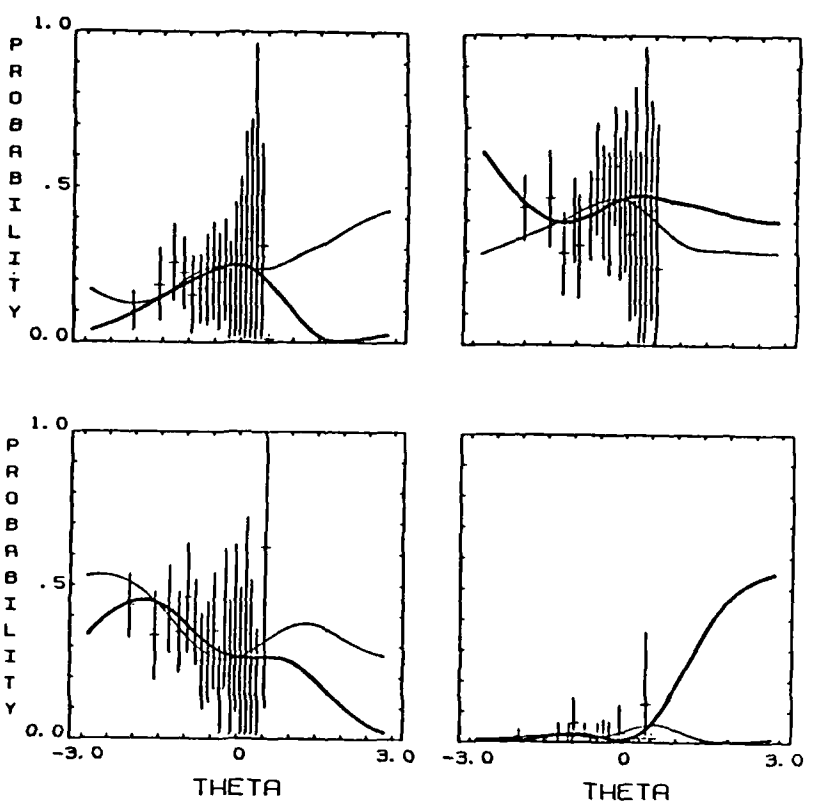
APPENDIX B: ESTIMATED COCCS, SIMULATION COCCS, AND EMPIRICAL PROPORTIONS FROM ESTIMATION SAMPLE


ITEM 1
ITEM 2
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

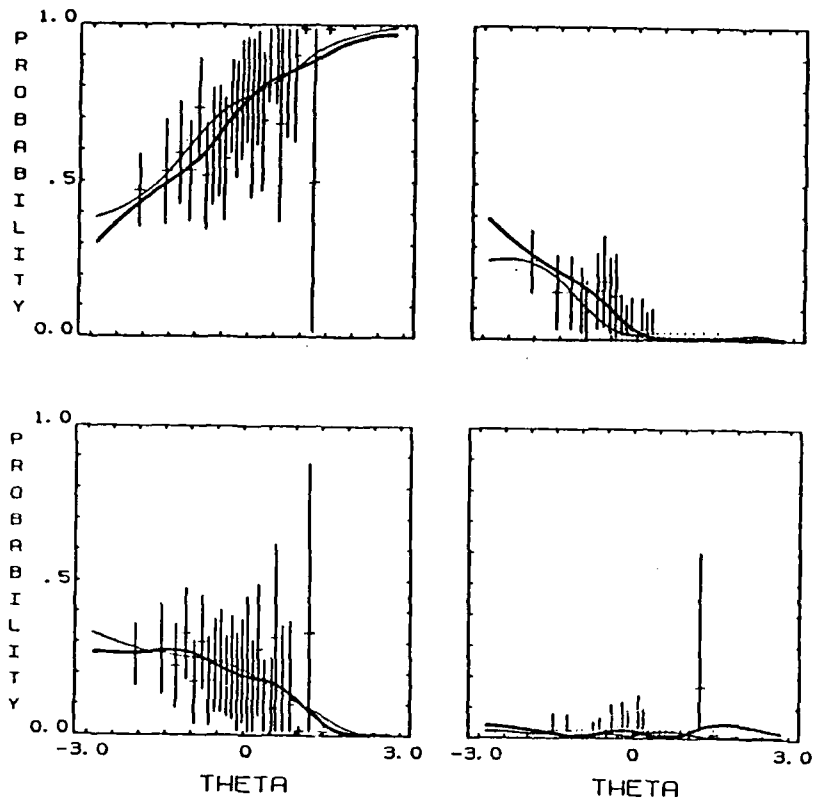
ITEM 3


1.0
PROBABILITY
.5
0.0
-3.0
0
3.0
THETA

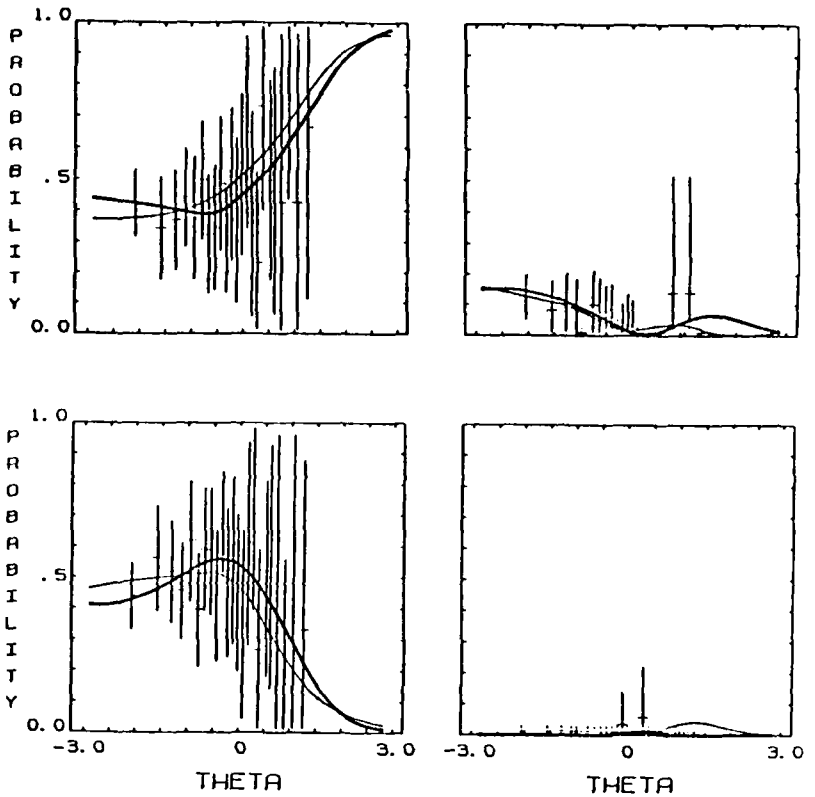
ITEM 4

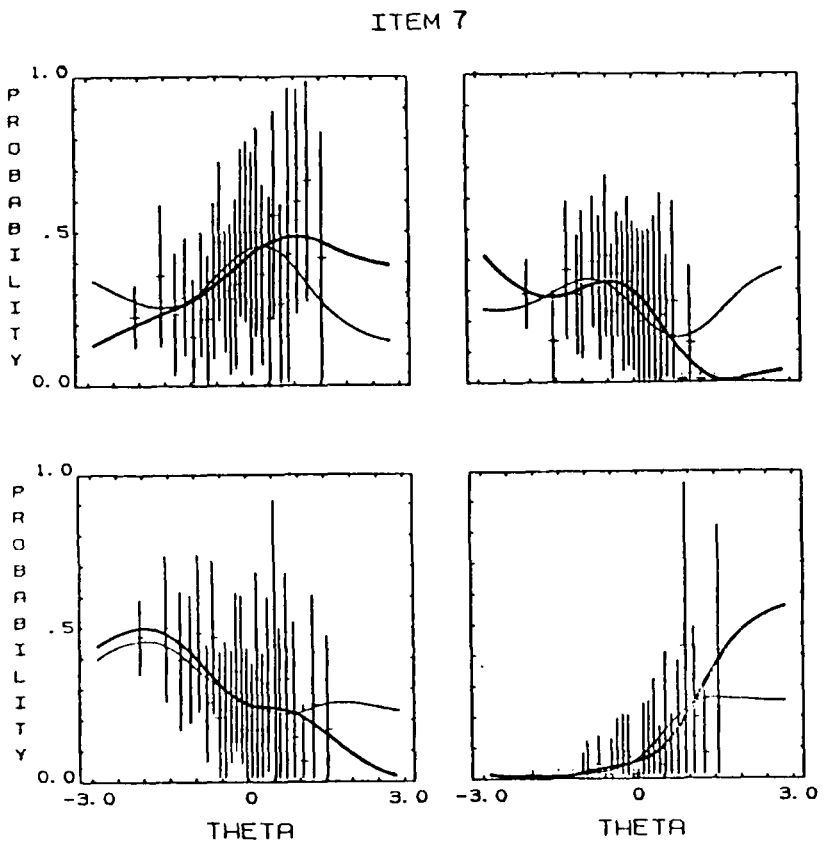

1.0
PROBABILITY
.5
0.0
-3.0
0
3.0
THETA

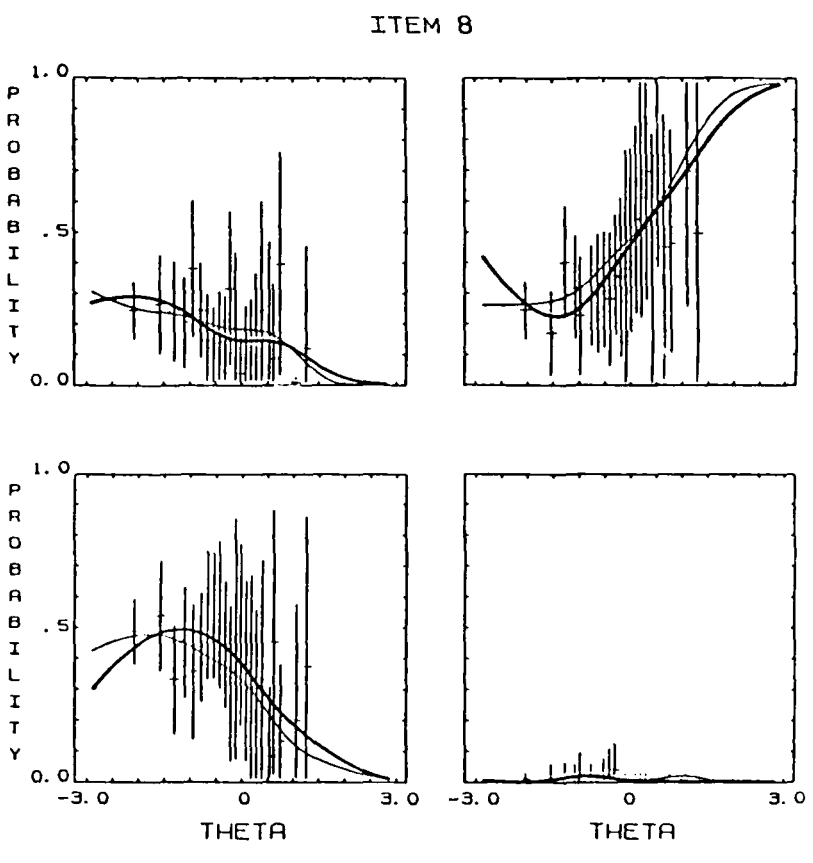
ITEM 5

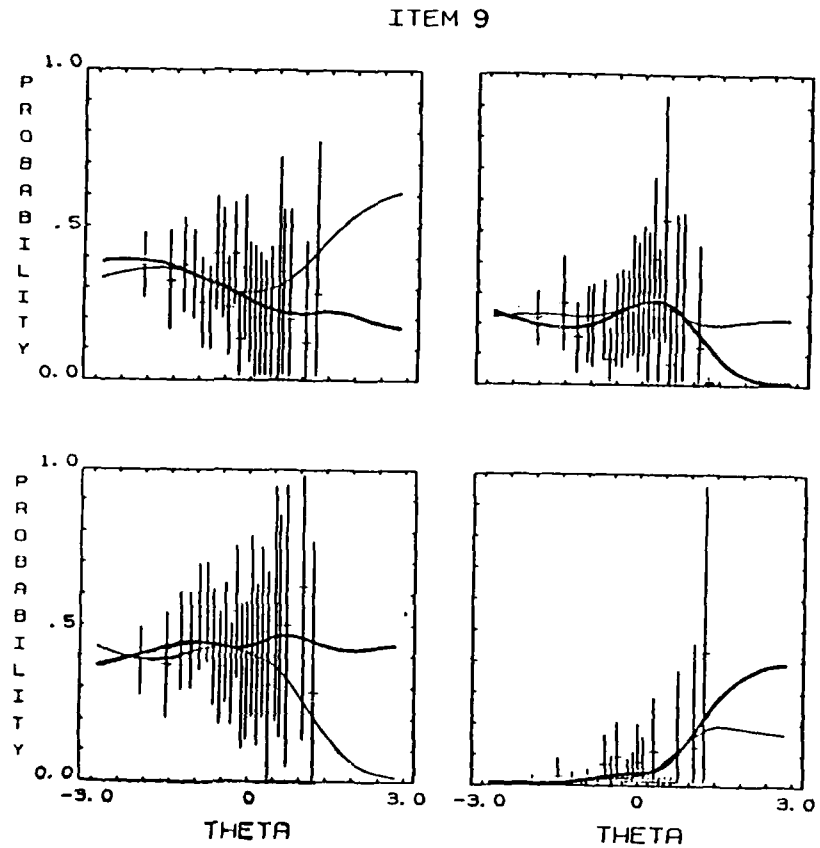

1.0
PROBABILITY
.5
0.0
-3.0
0
3.0
THETA

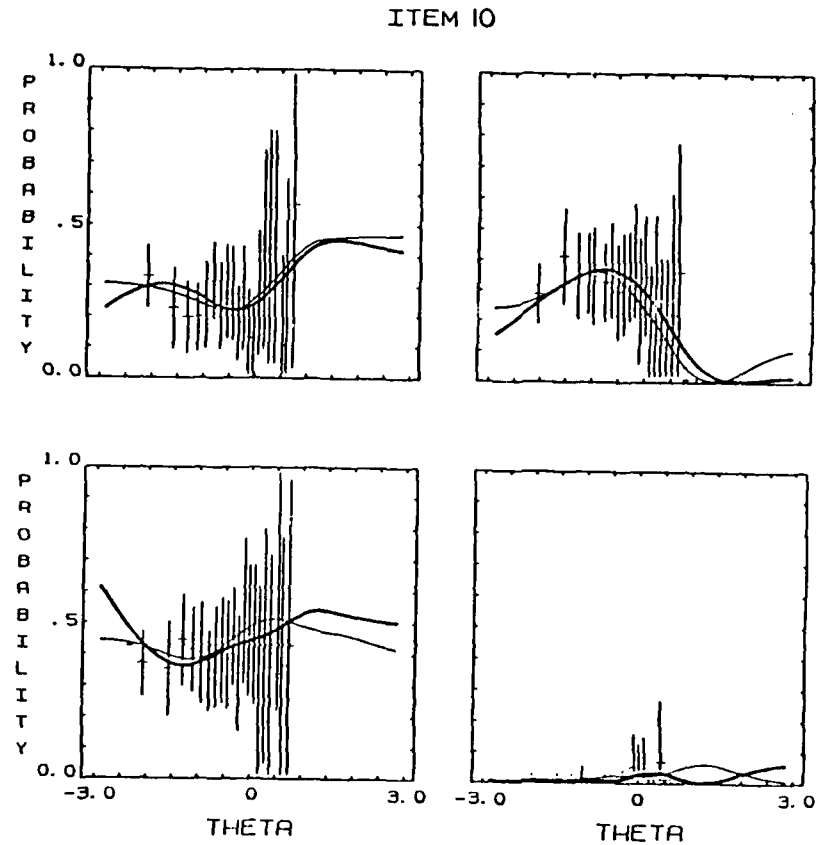
ITEM 6

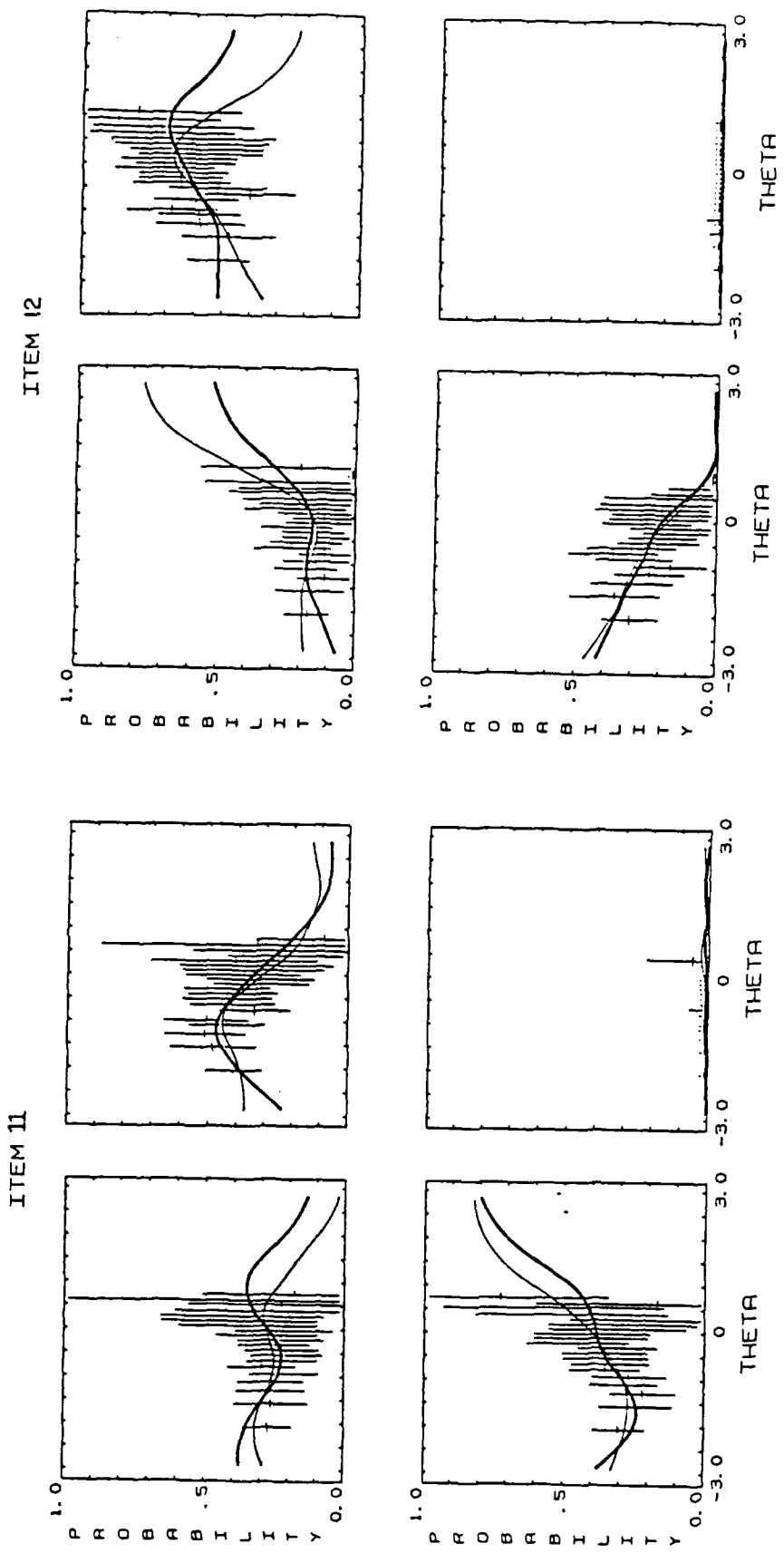

1.0
PROBABILITY
.5
0.0
-3.0
0
3.0
THETA

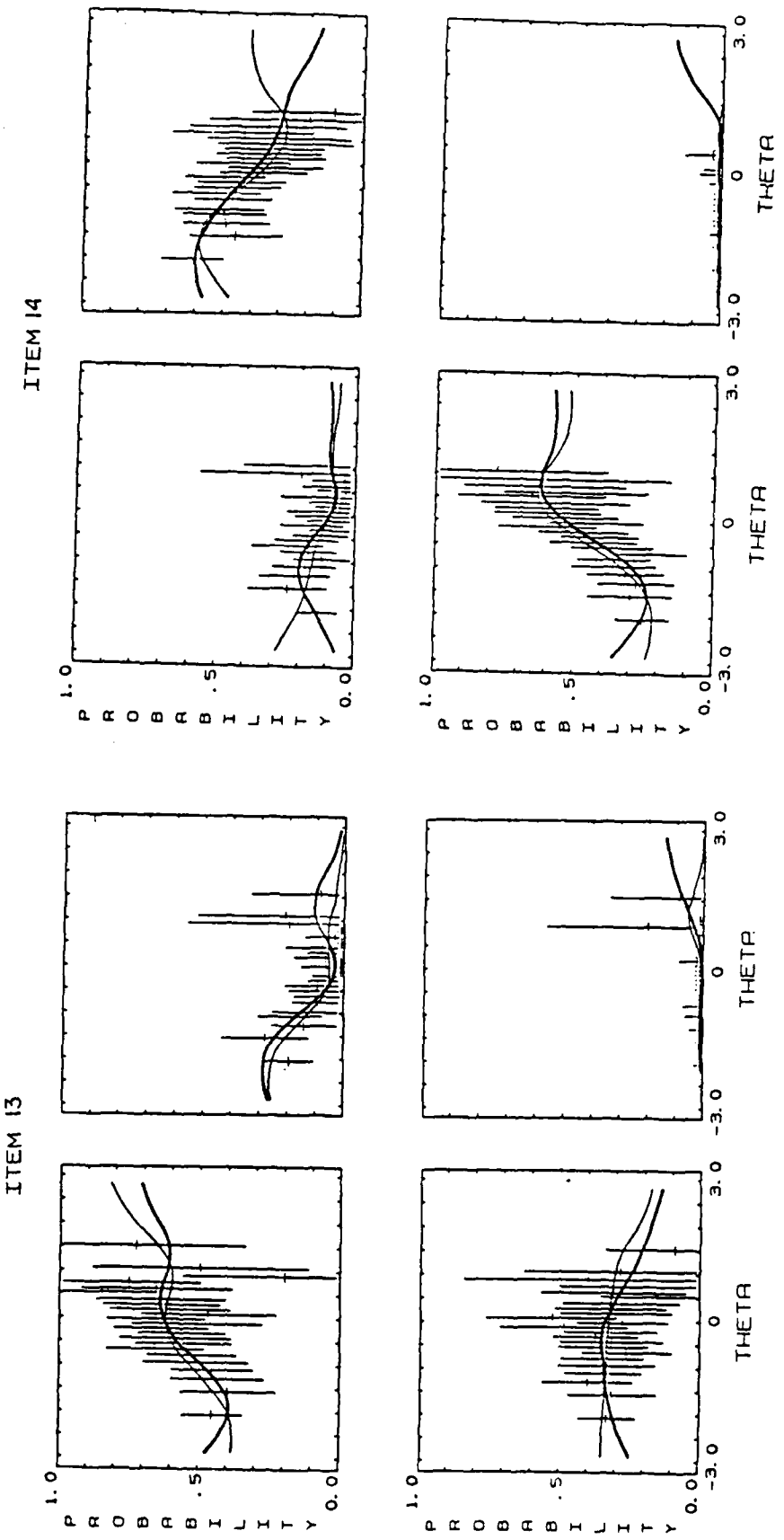

ITEM 7
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

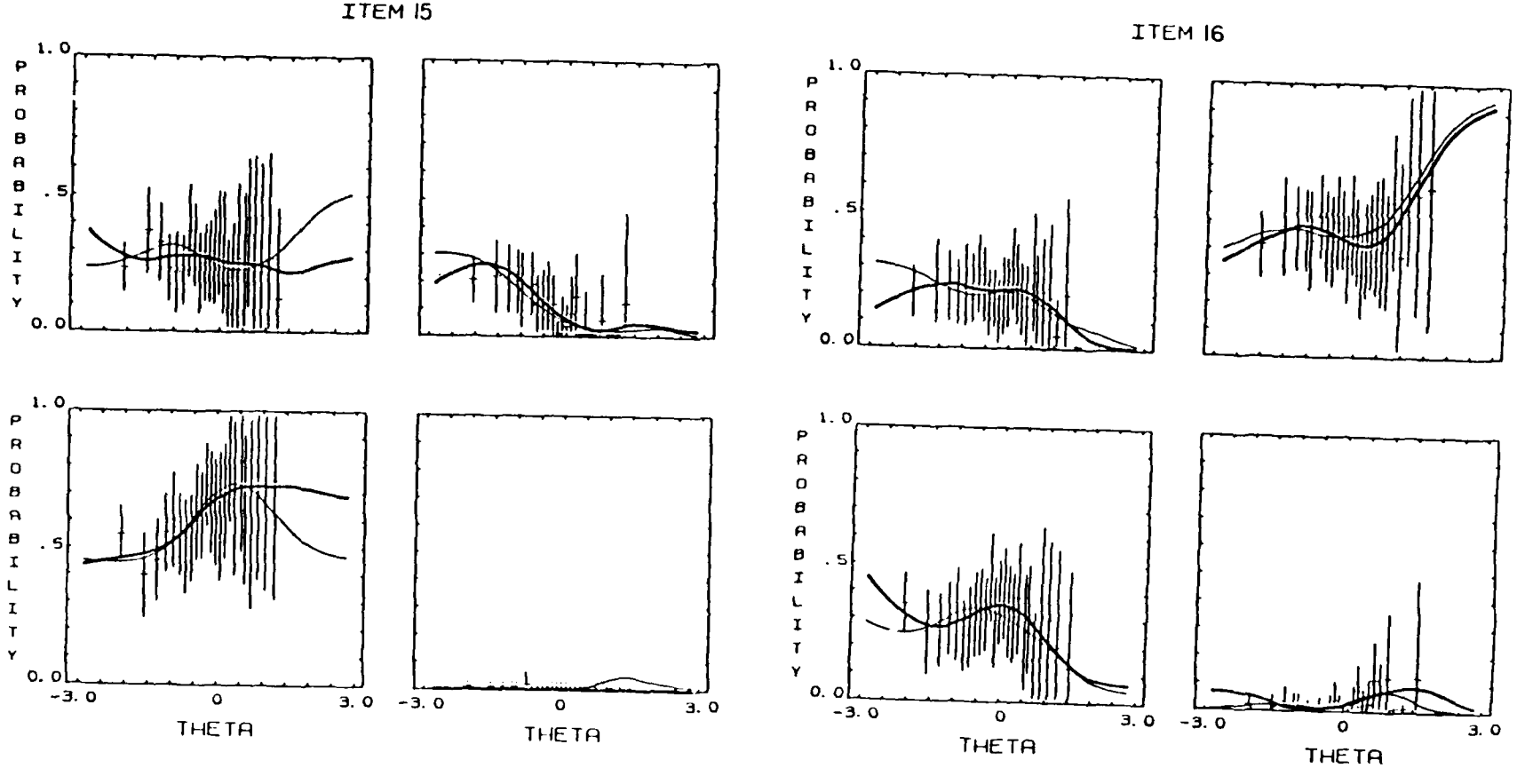

ITEM 8
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

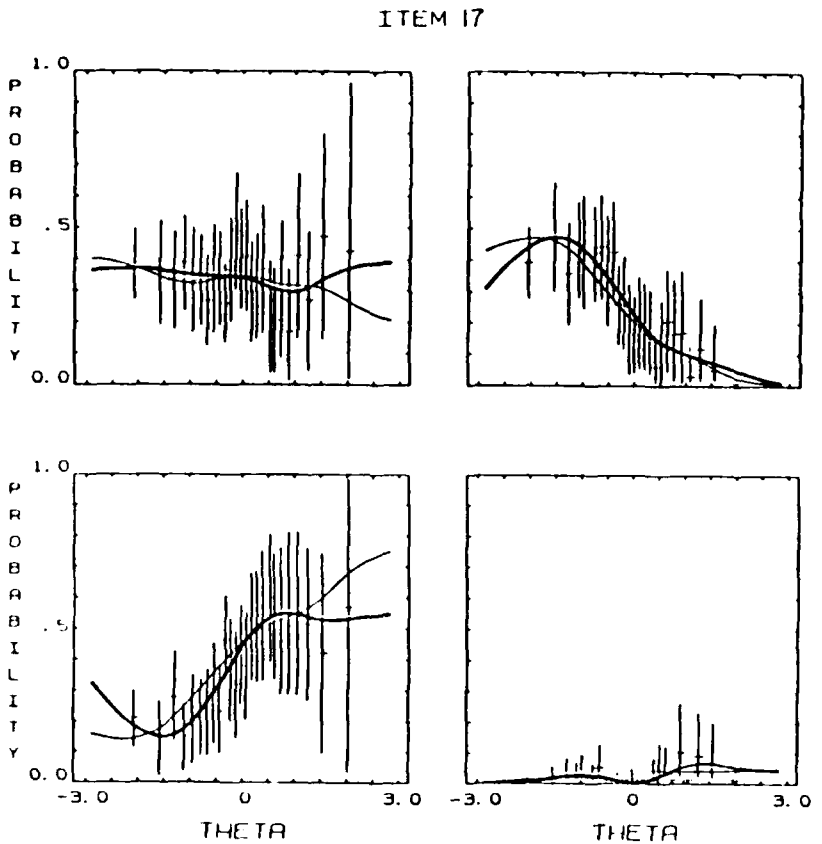

ITEM 9
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

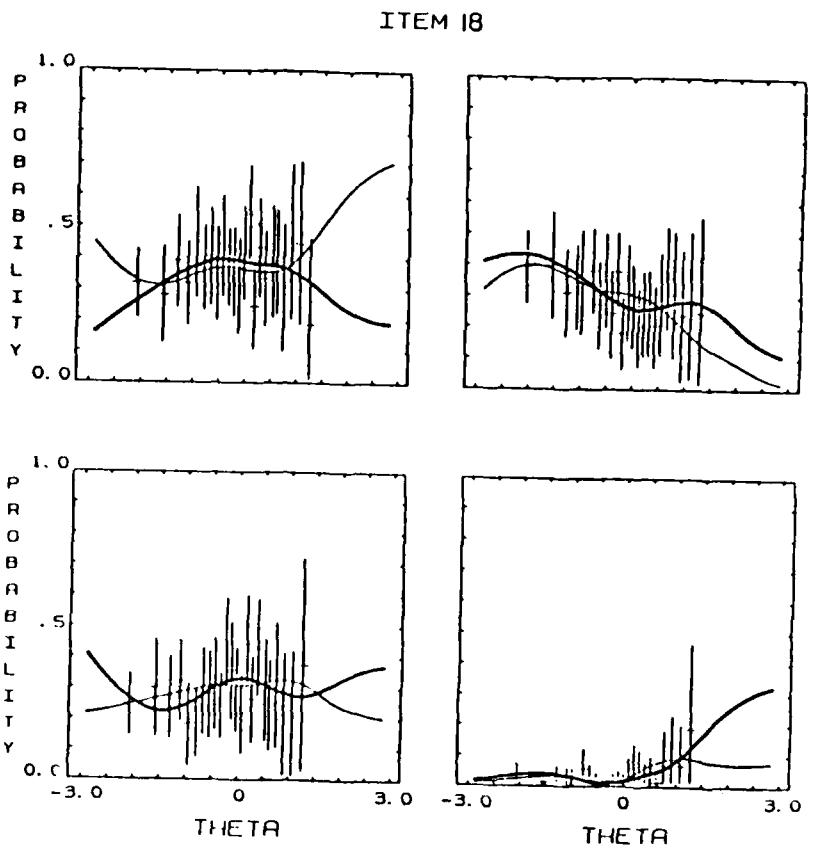

ITEM 10
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

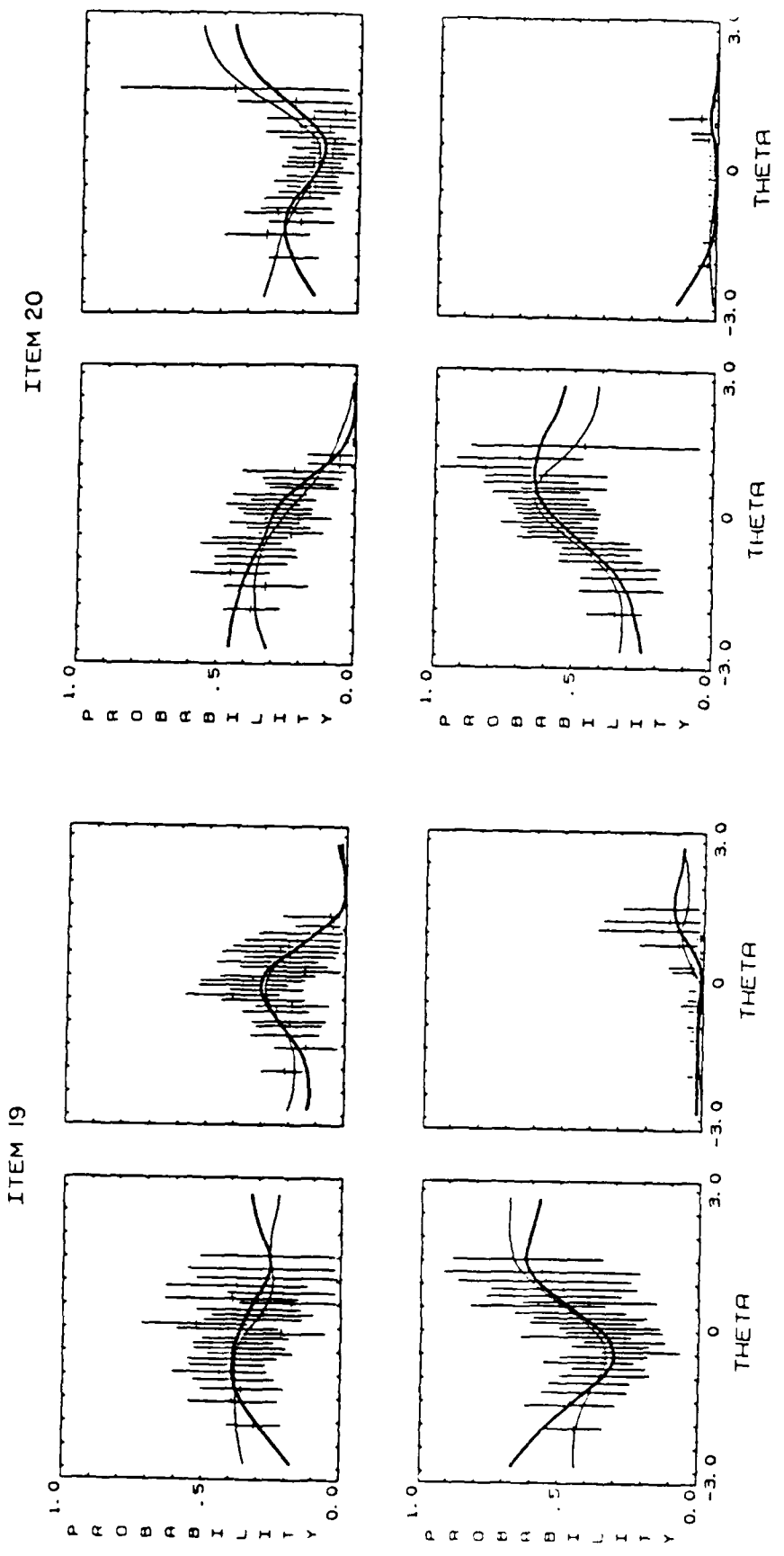

ITEM 11
ITEM 12
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

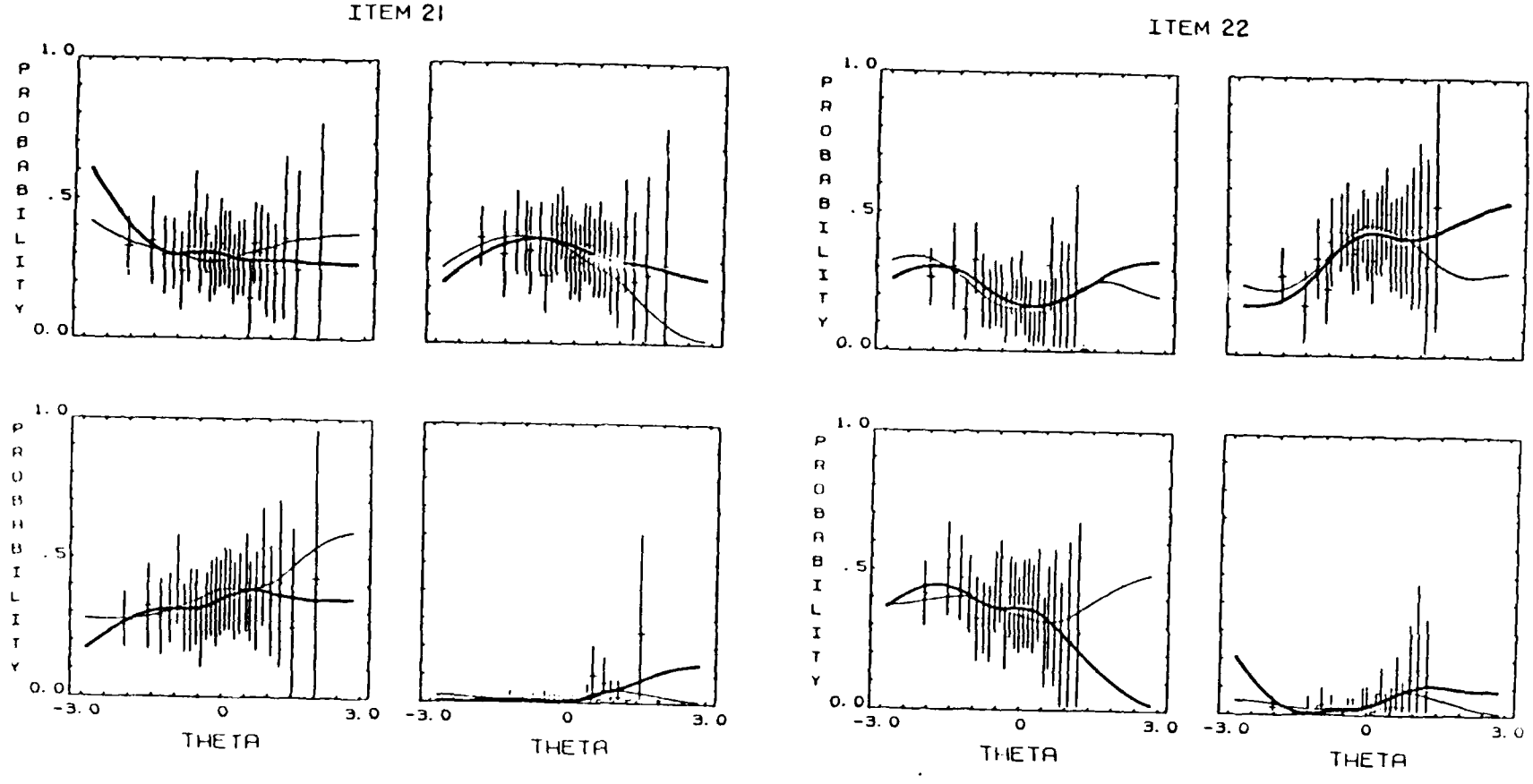

ITEM 13
ITEM 14
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

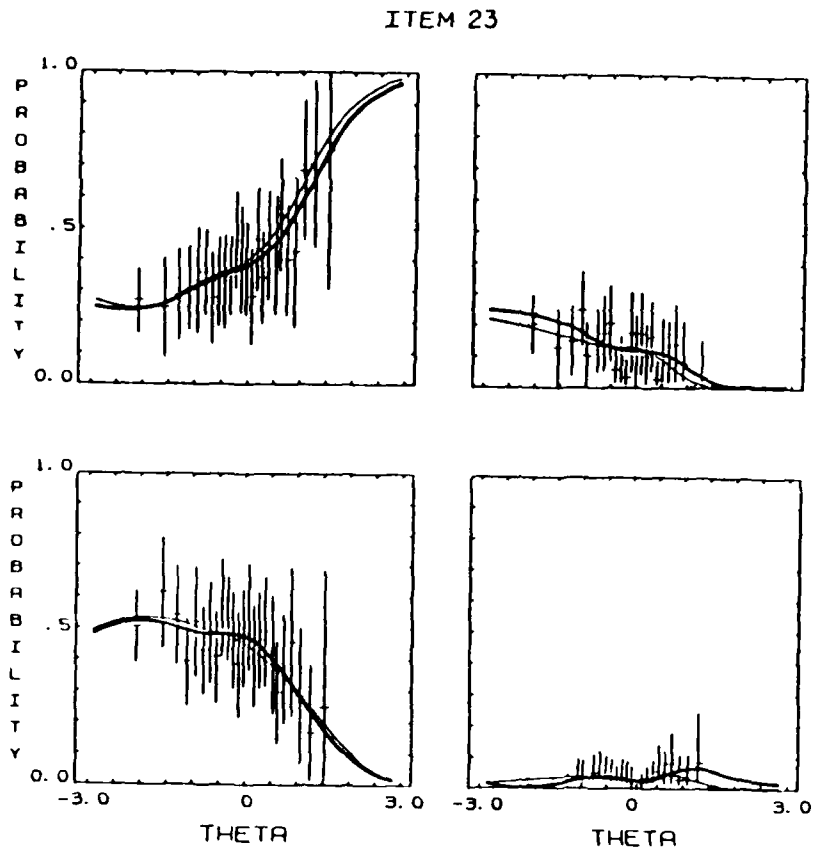

ITEM 15
ITEM 16
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

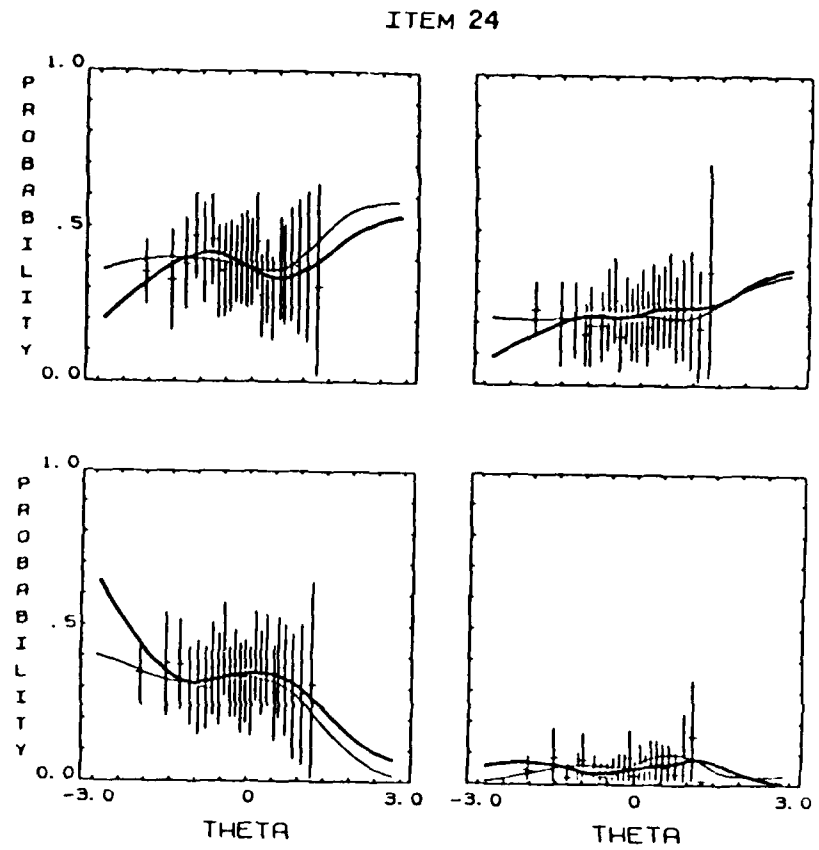

ITEM 17
1.0
PROBABILITY
.5
0.0
-3.0
0
3.0
THETA

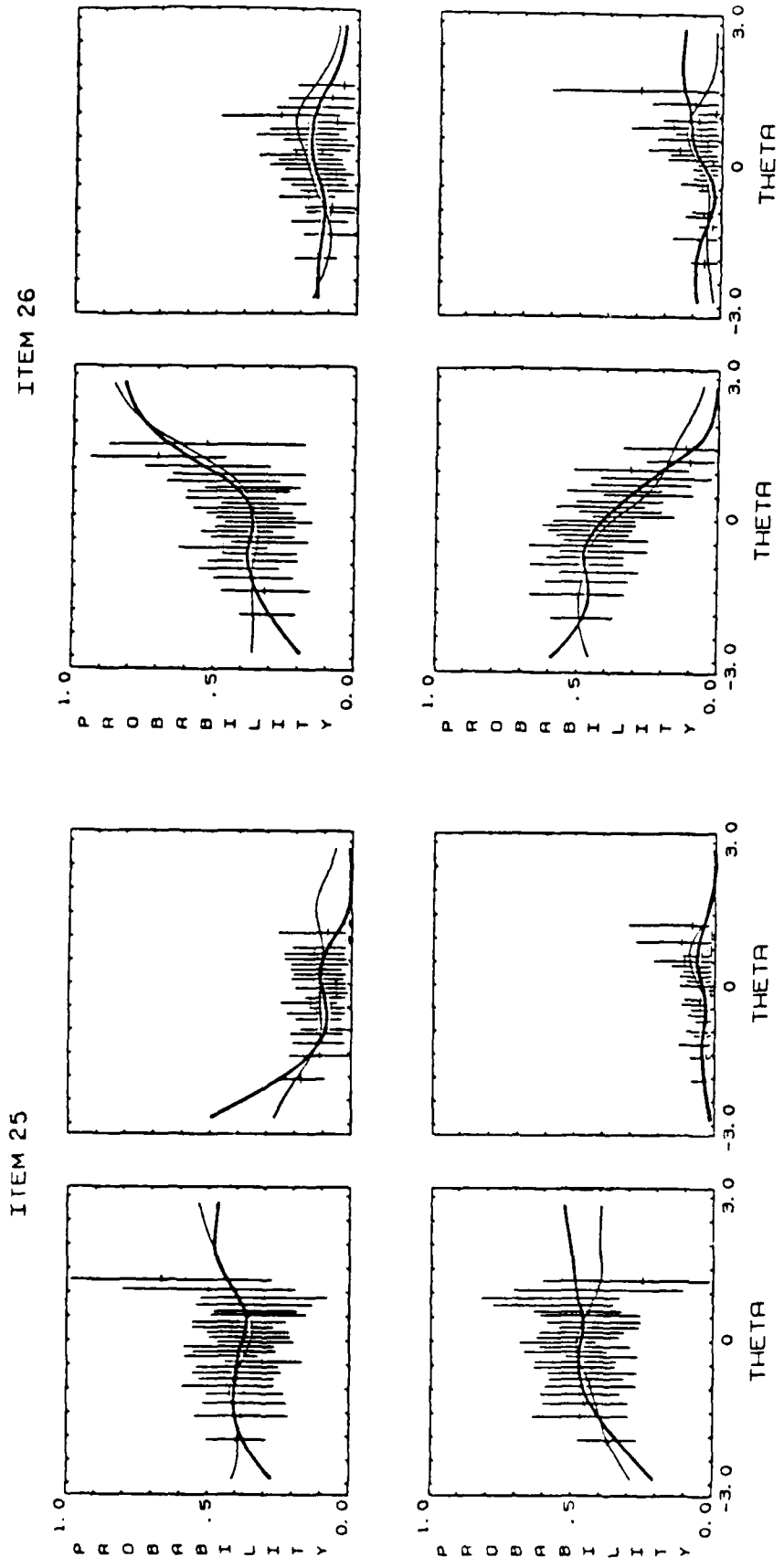

ITEM 18
1.0
PROBABILITY
.5
0.0
-3.0
0
3.0
THETA

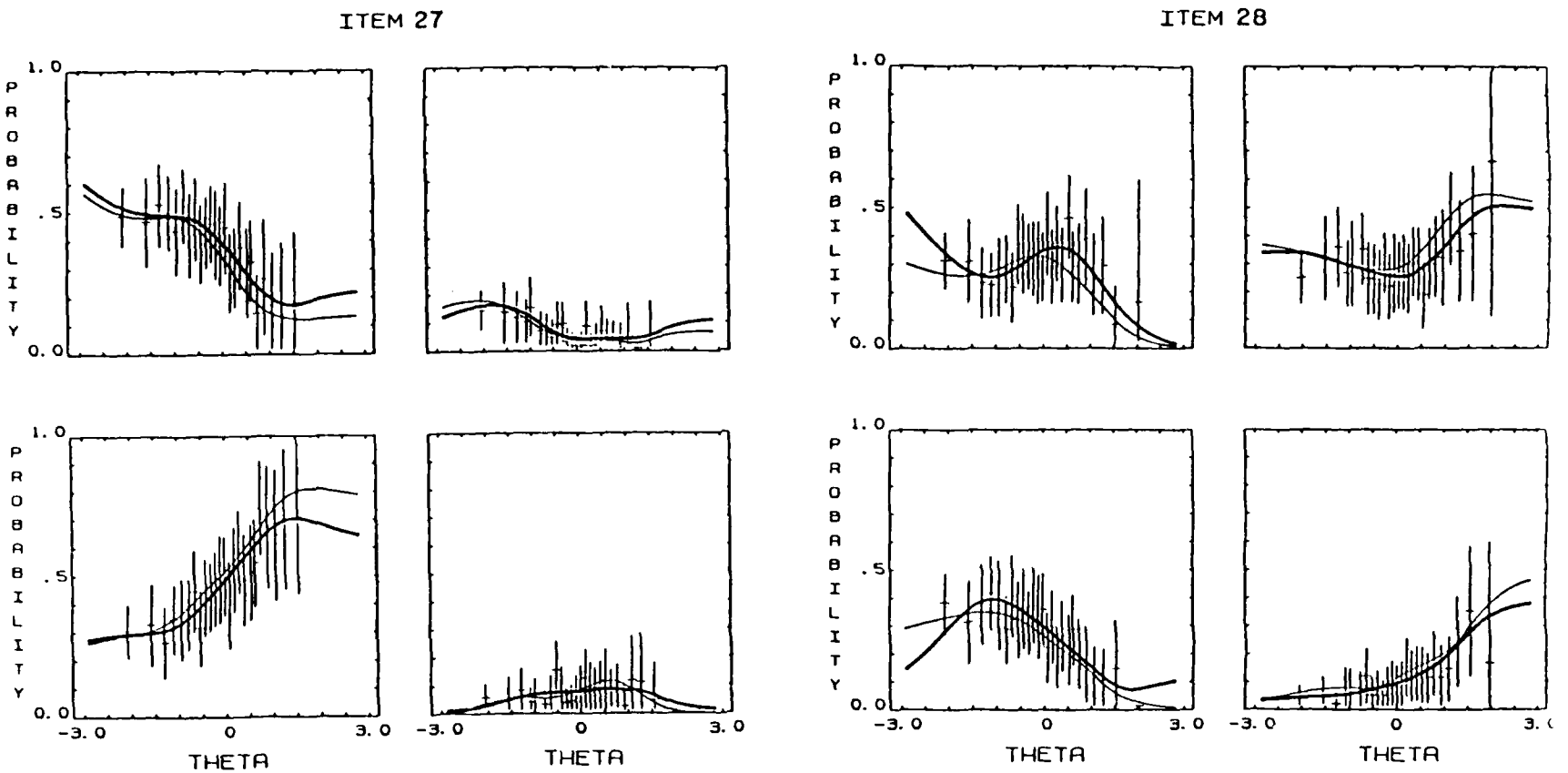

ITEM 19
ITEM 20
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA


ITEM 21
ITEM 22
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

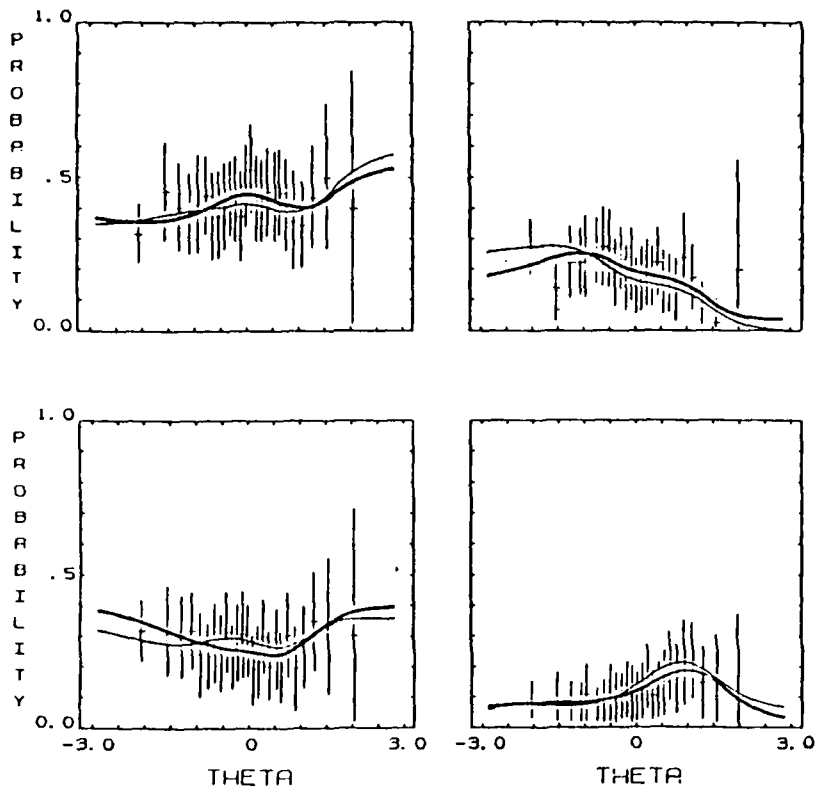

ITEM 23
1.0
PROBABILITY
.5
0.0
-3.0
0
3.0
THETA


ITEM 24
1.0
PROBABILITY
.5
0.0
-3.0
0
3.0
THETA

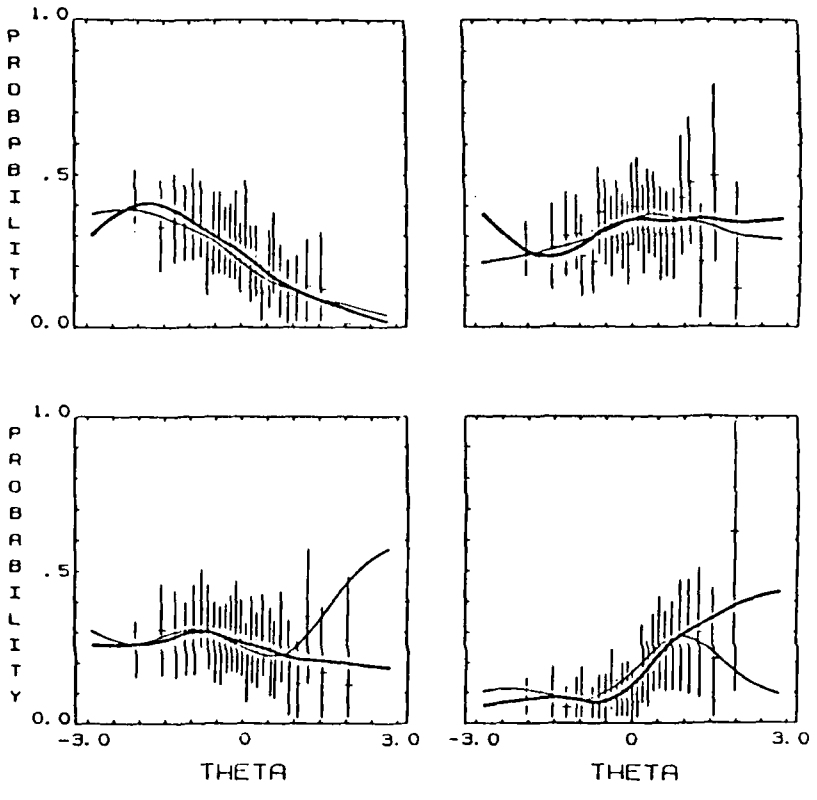

ITEM 25
ITEM 26
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA


ITEM 27
ITEM 28
PROBABILITY
1.0
.5
0.0
-3.0
0
3.0
THETA

ITEM 29



ITEM 30



APPENDIX C: MULTITEST EXTENSIONS OF OPTIMAL INDICES

An approximation to the likelihoods required for an optimal statistic for two unidimensional tests is given in this appendix. The approach easily generalizes to $\underline{m} > 2$ dimensions.

To begin, rewrite $\underline{F}^*$ from Equation 31 as

$$\underline{F}^* = \int\int \{P(\mathbf{U}_1 = u_1|\theta_1)[\phi(\theta_1)/\phi(\theta_1)]\}$$

$$\cdot\{P(\mathbf{U}_2 = u_2|\theta_2)[\phi(\theta_2)/\phi(\theta_2)]\}\ \phi_2(\boldsymbol{\theta};\mathbf{0},\boldsymbol{\Sigma})d\boldsymbol{\theta}$$

$$\doteq \int\int [e^{a_1\theta_1^2+b_1\theta_1+c_1}/\phi(\theta_1)]\ [e^{a_2\theta_2^2+b_2\theta_2+c_2}/\phi(\theta_2)]\cdot\phi_2(\boldsymbol{\theta};\mathbf{0},\boldsymbol{\Sigma})d\boldsymbol{\theta}$$

$$= \int\int e^{a_1\theta_1^2+b_1\theta_1+c_1}\ e^{(1/2)\theta_1^2}\ e^{a_2\theta_2^2+b_2\theta_2+c_2}\ e^{(1/2)\theta_2^2}$$

$$\cdot(\det\boldsymbol{\Sigma})^{-1/2}\ e^{(\theta_1^2-2\rho\theta_1\theta_2+\theta_2^2)/2(\rho^2-1)}d\boldsymbol{\theta}$$

$$= \underline{F}\ ,$$

where $\phi(\cdot)$ is the standard normal density. For the next step in our analysis, it is useful to rewrite this equation in matrix notation. Consequently, let

$$\mathbf{A}_1 = \begin{bmatrix} a_1 & 0 \\ 0 & 0 \end{bmatrix}\ , \qquad \mathbf{b}_1 = \begin{bmatrix} b_1 \\ 0 \end{bmatrix}\ ,$$

$$\mathbf{A}_2 = \begin{bmatrix} 0 & 0 \\ 0 & a_2 \end{bmatrix}\ , \qquad \mathbf{b}_2 = \begin{bmatrix} 0 \\ b_2 \end{bmatrix}\ ,$$

$$\mathbf{K}_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \end{bmatrix}\ , \qquad \mathbf{K}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1/2 \end{bmatrix}\ ,$$

$$\mathbf{K}_3 = \frac{1}{2(\rho^2-1)}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}$$

Then

$$\underline{F} = (\det\boldsymbol{\Sigma})^{-1/2}\int\int \exp[\boldsymbol{\theta}'\mathbf{A}_1\boldsymbol{\theta} - \mathbf{b}_1'\boldsymbol{\theta} + \underline{c}_1 + \boldsymbol{\theta}'\mathbf{K}_1\boldsymbol{\theta}$$

$$+ \boldsymbol{\theta}'\mathbf{A}_2\boldsymbol{\theta} + \mathbf{b}_2'\boldsymbol{\theta} + \underline{c}_2 + \boldsymbol{\theta}'\mathbf{K}_2\boldsymbol{\theta} + \boldsymbol{\theta}'\mathbf{K}_3\boldsymbol{\theta}]d\boldsymbol{\theta}$$

$$= (\det\boldsymbol{\Sigma})^{-1/2}\ e^{c_1+c_2}\int\int \exp[\boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta} + \mathbf{b}'\boldsymbol{\theta}]d\boldsymbol{\theta}\ ,$$

where

$$\mathbf{A} = \mathbf{A}_1 + \mathbf{K}_1 + \mathbf{A}_2 + \mathbf{K}_2 + \mathbf{K}_3$$

and

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2 \ .$$

To complete the square in the exponent of the above integrand, notice that since $\mathbf{A}$ is symmetric,

$$\theta'\mathbf{A}\theta + \mathbf{b}'\mathbf{A}^{-1}\mathbf{A}\theta + \tfrac{1}{4}\mathbf{b}'\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{b} - \tfrac{1}{4}\mathbf{b}'\mathbf{A}^{-1}\mathbf{b}$$

$$= (\theta + \tfrac{1}{2}\mathbf{A}^{-1}\mathbf{b})'\mathbf{A}(\theta + \tfrac{1}{2}\mathbf{A}^{-1}\mathbf{b}) - \tfrac{1}{4}\mathbf{b}'\mathbf{A}^{-1}\mathbf{b} \ ,$$

provided that $\mathbf{A}$ is negative definite. Diagonalize $\mathbf{A}$ by $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}'$, where $\mathbf{V}'\mathbf{V} = \mathbf{I}$, let $\underline{k} = -\frac{1}{4}\mathbf{b}'\mathbf{A}^{-1}\mathbf{b}$, and let $\underline{w} = (\det\Sigma)^{-1/2}e^{c_1+c_2+k}$. Then

$$\underline{F} = (\det\Sigma)^{-1/2}\, e^{c_1+c_2+k} \int\int \exp[(\theta + \tfrac{1}{2}\mathbf{A}^{-1}\mathbf{b})'\ \mathbf{V}\Lambda\mathbf{V}'\ (\theta + \tfrac{1}{2}\mathbf{A}^{-1}\mathbf{b})]d\theta$$

$$= \underline{w} \int\int \exp[\theta'\mathbf{V}\Lambda\mathbf{V}'\theta]d\theta$$

$$= \underline{w} \int\int \exp[\mathbf{t}'\Lambda\mathbf{t}]d\mathbf{t} \ ,$$

where $\mathbf{t} = (t_1, t_2)' = \theta'\mathbf{V}$, because the Jacobian of the transformation is one. The middle equality above holds because the volume of the bivariate density is unaffected by the location parameter. Since Λ is diagonal with negative diagonal elements λ_1 and λ_2,

$$\underline{F} = \underline{w} \int \exp[-\tfrac{1}{2}\,\underline{t}_1^2(-2\lambda_1)]dt_1 \int \exp[-\tfrac{1}{2}\,\underline{t}_2^2(-2\lambda_2)]dt_2$$

$$= \underline{w} \int \exp[-\tfrac{1}{2}\,\underline{t}_1^2 \,/\, \sigma_1^2]dt_1 \int \exp[-\tfrac{1}{2}\,\underline{t}_2^2 \,/\, \sigma_2^2]dt_2$$

$$= 2\pi w\sigma_1\sigma_2 \ ,$$

where $-2\lambda_j = 1/\sigma_j^2$, $j = 1,2$. Because $2\sigma_1\sigma_2 = 1/\sqrt{\lambda_1\lambda_2} = (\det\Lambda)^{-1/2}$ $= (\det\mathbf{A})^{-1/2}$, we obtain

$$\underline{F} = \pi\underline{w}(\det\mathbf{A})^{-1/2}$$

$$= \pi \exp[\underline{c}_1 + c_2 - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}/4]\ (\det\Sigma)^{-1/2}\ (\det\mathbf{A})^{-1/2}$$

as the final expression for our approximation to $\underline{F}^*$ given in Equation 31.